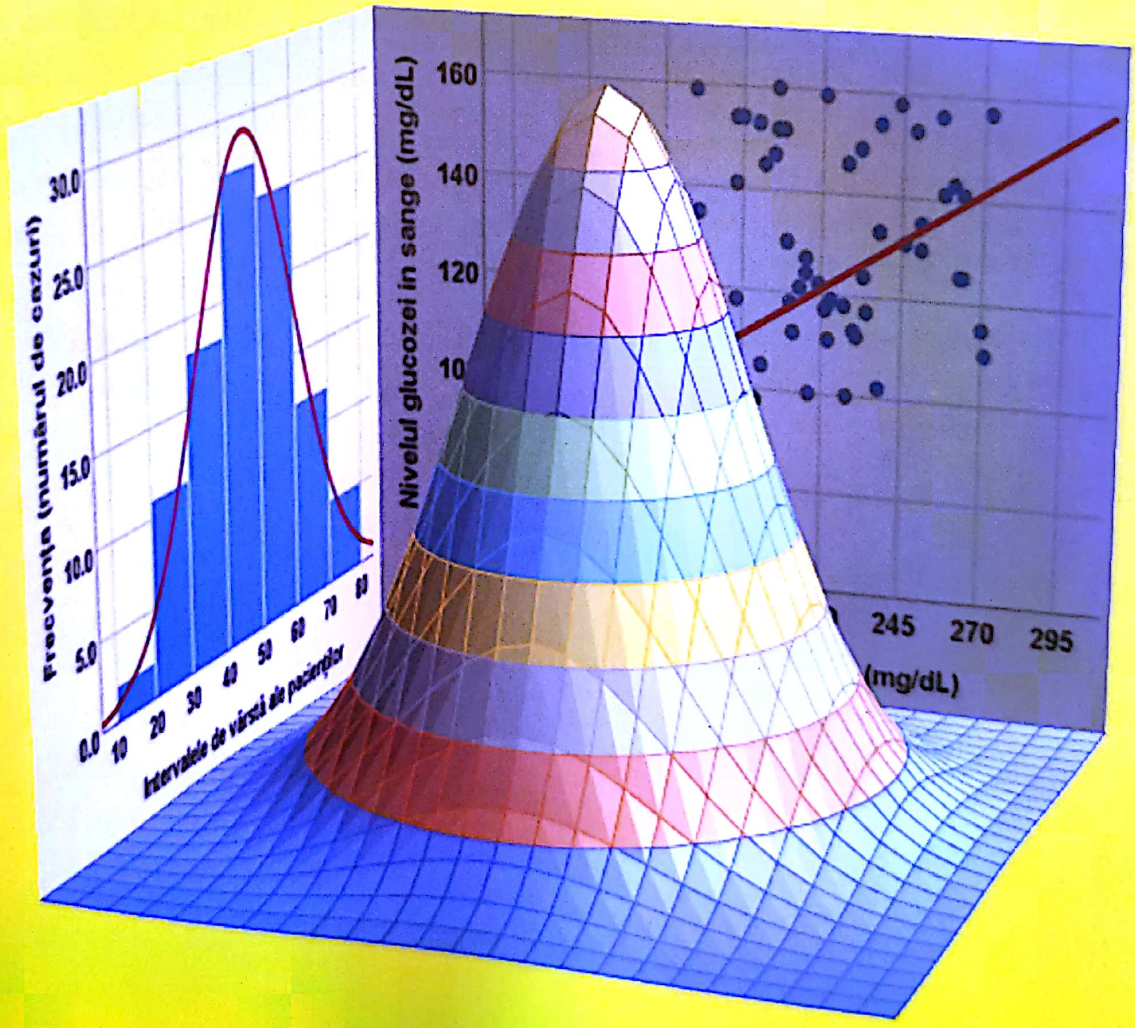
Omul și societatea contemporană

Viorel ȚARCĂ

NOȚIUNI FUNDAMENTALE DE BIOSTATISTICĂ



JUNIMEA

Scanned with OKEN Scanner

Viorel ŢARCĂ este licenţiat al Universităţii „Alexandru Ioan Cuza" din Iaşi, Facultatea de Ştiinţe Economice, Secţia Statistică Economico-Socială.

Studii aprofundate în specializarea Statistică Social-Economică, Facultatea de Ştiinţe Economice, Universitatea „Alexandru Ioan Cuza", Iaşi

Doctor în Economie, Specialitatea Relaţii Economice Internaţionale, Facultatea de Economie şi Administrarea Afacerilor, Universitatea „Alexandru Ioan Cuza", Iaşi.

În prezent este conferenţiar doctor la Universitatea „Apollonia" din Iaşi.

Scanned with OKEN Scanner

Coordonatorii colecției:
Ioan NECULA
Radu CUCUTEANU
Ioan-Alexandru TOFAN

Descrierea CIP a Bibliotecii Naționale a României
ȚARCĂ, VIOREL
NOȚIUNI FUNDAMENTALE DE BIOSTATISTICĂ / Viorel ȚARCĂ
Iași: Junimea, 2025

ISBN 978-973-37-2931-0

316

Redactor: Frăguța ZAHARIA
Coperta: Ionuț BROȘTIANU
Tehnoredactor: Cezar BACIU

Editura JUNIMEA,
Bulevardul Carol I
Parcul Copou, Iași
700 506

C. P. 85, Oficiul Poștal nr. 1, Iași
e-mail: edjunimea@gmail.com
revistascriptor@gmail.com
Telefon redacție: 0232 705 837

© Viorel ȚARCĂ
© Editura JUNIMEA, IAȘI – ROMÂNIA

Scanned with OKEN Scanner

Scanned with OKEN Scanner

Viorel ȚARCĂ

NOȚIUNI FUNDAMENTALE DE BIOSTATISTICĂ

EDITURA JUNIMEA
IAȘI, 2025

CUPRINS

1. NOȚIUNILE DE BAZĂ ALE STATISTICII

1.1. Introducere 9
1.2. Definiția și importanța practică a statisticii 11
1.3. Scurt istoric al statisticii românești 15
1.4. Tipuri de analiză statistică – descriptivă și inferențială 17
1.5. Categorii de variabile utilizate în analiza statistică 19
1.6. Nivelurile de măsurare a datelor 21
1.7. Variabilitatea datelor medicale 25
1.8. Erorile de înregistrare 26

2. PREZENTAREA INFORMAȚIILOR STATISTICE

2.1. Procesul de centralizare și grupare a informațiilor statistice 28
2.2. Construirea distribuțiilor de frecvențe și a graficelor statistice în cazul variabilelor cantitative - histograma și poligonul frecvențelor 30
2.3. Tabelele de frecvențe și reprezentarea grafică a variabilelor calitative - diagramele prin bare și graficele de structură 37
2.4. Prezentarea relației sau asocierii dintre două variabile cantitative - diagrama tip „nor de puncte" 39
2.5. Prezentarea relației sau asocierii dintre două variabile calitative - tabelele de contingență și graficul prin coloane .. 41
2.6. Prezentarea relației sau asocierii dintre o variabilă cantitativă și una calitativă - piramida vârstelor 43

3. INDICATORII TENDINȚEI CENTRALE, VARIAȚIEI ȘI FORMEI DISTRIBUȚIILOR

3.1. Indicatorii tendinței centrale 45
3.1.1. Media aritmetică 47
3.1.2. Medii cu aplicații speciale 50
3.1.3. Mediana 53
3.1.4. Modul 55
3.2. Indicatorii variației sau dispersiei 57
3.2.1. Amplitudinea variației 58
3.2.2. Varianța și deviația standard 58
3.2.3. Coeficientul de variație 60
3.2.4. Cuartilele, decilele și centilele 61

Scanned with OKEN Scanner

3.3. Indicatorii formei distribuțiilor 65
3.3.1. Coeficientul de asimetrie 66
3.3.2. Coeficientul de boltire 68
3.4. Importanța practică a deviației standard 70
3.4.1. Scorurile z 70
3.4.2. Teorema lui Cebîșev și regula empirică 72

4. CONCEPTE DE BAZĂ ALE TEORIEI PROBABILITĂȚILOR
4.1. Definiții și abordări ale conceptului de probabilitate 74
4.2. Regulile adunării probabilităților 78
4.3. Regulile înmulțirii probabilităților 80
4.4. Principiile numărării 82
4.5. Formula probabilității totale 84
4.6. Teorema lui Bayes 85

5. DISTRIBUȚII PROBABILISTICE DISCRETE ȘI CONTINUE
5.1. Definiția unei distribuții probabilistice 87
5.2. Variabilele aleatoare discrete și continue 90
5.3. Media aritmetică și varianța unei distribuții probabilistice discrete 94
5.4. Media aritmetică și varianța unei distribuții probabilistice continue 95
5.5. Distribuții probabilistice discrete - binomială, hipergeometrică, Poisson 96
5.6. Distribuții probabilistice continue – uniformă și normală 110
5.7. Distribuția normală standard 115

6. METODE DE EȘANTIONARE ȘI TEOREMA LIMITĂ CENTRALĂ
6.1. Avantajele eșantionării 120
6.2. Principalele metode de eșantionare 122
6.3. Distribuțiile de sondaj și eroarea de sondaj 125
6.4. Distribuția de sondaj a mediilor eșantioanelor 126
6.5. Distribuția de sondaj a proporțiilor eșantioanelor 133
6.6. Teorema limită centrală 137
6.7. Aplicații practice ale distribuției de sondaj a mediilor eșantioanelor 139

7. ESTIMAREA ȘI INTERVALELE DE ÎNCREDERE
7.1. Estimarea punctuală și prin interval de încredere a parametrilor unei populații 144

Scanned with OKEN Scanner

7.2. Intervalul de încredere pentru estimarea mediei unei populații 147
7.3. Intervalul de încredere pentru estimarea diferenței dintre mediile a două populații 157
7.4. Intervalul de încredere pentru estimarea proporției unei populații 163
7.5. Intervalul de încredere pentru estimarea diferenței dintre proporțiile a două populații 166

8. TESTAREA IPOTEZELOR STATISTICE. VOLUMUL OPTIM AL EȘANTIONULUI
8.1. Definiția și principalele caracteristici ale unui test statistic 169
8.2. Erorile specifice procesului de testare a ipotezelor statistice 176
8.3. Determinarea volumului eșantionului de studiu pentru estimarea mediei unei populații 179
8.4. Determinarea volumului eșantionului de studiu pentru estimarea proporției unei populații 184

9. TESTELE PARAMETRICE DE VERIFICARE A IPOTEZELOR STATISTICE
9.1. Definiția, etapele și scopul unui test parametric de verificare a ipotezelor statistice 188
9.2. Testarea ipotezelor pentru media unei populații - testele „z" și „t" 192
9.3. Testarea ipotezelor pentru proporția unei populații - testul „z" 204
9.4. Testarea ipotezelor pentru diferența mediilor a două populații - testele „z" și „t" 207
9.5. Testarea ipotezelor pentru diferența proporțiilor a două populații - testul „z" 225
9.6. Testarea ipotezelor pentru raportul variațelor a două populații - testul „F" 229

10. ANALIZA VARIANȚEI - ANOVA
10.1. Definiția și principalele tipuri ale analizei variației 234
10.2. Prezentarea modelul ANOVA unifactorială 236
10.3. Condițiile preliminare necesare în vederea aplicării testului ANOVA unifactorială 240
10.4. Componentele principale ale variabilității populației generale - tabelul de calcul ANOVA 241

Scanned with OKEN Scanner

10.5. Testele de comparații multiple – analiza POST-HOC 254

11. CORELAȚIE ȘI REGRESIE SIMPLĂ LINIARĂ

11.1. Corelația simplă liniară – noțiuni preliminare 258
11.2. Coeficientul de corelație simplă liniară 262
11.3. Testarea semnificației coeficientului de corelație 268
11.4. Analiza de regresie simplă liniară - ecuația de regresie și metoda celor mai mici pătrate 271
11.5. Testarea semnificației pantei dreptei de regresie 275
11.6. Testarea modelului de regresie simplă liniară prin metoda ANOVA 278
11.7. Coeficientul de determinare simplă liniară 281
11.8. Presupunerile modelului de regresie simplă liniară 283

12. TESTAREA ASOCIERII DINTRE VARIABILELE CATEGORIALE

12.1. Testul CHI-pătrat și formele lui 289
12.2. Testul CHI-pătrat de evaluare a independenței 293
12.3. Testul CHI-pătrat de evaluare a omogenității 301
12.4. Riscul relativ utilizat în cadrul studiilor prospective 304
12.5. Raportul cotelor utilizat în cadrul studiilor retrospective 309

ANEXE 313
BIBLIOGRAFIE 322

Scanned with OKEN Scanner

Capitolul. 1

NOȚIUNILE DE BAZĂ ALE BIOSTATISTICII

SUBIECTE:

1.1. Introducere
1.2. Definiția și importanța practică a statisticii
1.3. Scurt istoric al statisticii românești
1.4. Tipuri de analiză statistică – descriptivă și inferențială
1.5. Categorii de variabile utilizate în analiza statistică
1.6. Nivelurile de măsurare a datelor
1.7. Variabilitatea datelor medicale
1.8. Erorile de înregistrare

La finalizarea acestui capitol, studentul va fi capabil să:
O1-1: explice de ce cunoașterea noțiunilor statistice deține un rol important
O1-2: definească statistica și să ofere un exemplu practic de aplicare a acesteia
O1-3: facă diferența între statistica descriptivă și cea inferențială
O1-4: clasifice variabilele în calitative sau cantitative / discrete sau continue
O1-5: distingă între nivelurile de măsurare nominal, ordinal, interval și raport
O1-6: caracterizeze principalele tipuri de erori de înregistrare

1.1. INTRODUCERE

În lumea modernă a științei și medicinei, informațiile joacă un rol esențial în formularea ipotezelor, testarea teoriilor și luarea deciziilor bazate pe dovezi. Biostatistica, o ramură a statisticii aplicate în domeniul științelor biologice și medicale, oferă instrumentele necesare pentru colectarea, analiza și interpretarea acestor informații în mod riguros și obiectiv. Informațiile care prezintă interes în cadrul diferitelor studii medicale sunt reprezentate de „date", care ne sunt disponibile în special sub forma numerelor. În cadrul analizei statistice suntem interesați de două mari categorii de numere: cele obținute în urma realizării măsurătorilor propriu-

Scanned with OKEN Scanner

zise și cele care sunt rezultatul procesului de numărare. De exemplu, atunci când o asistentă medicală cântărește un pacient sau îi ia temperatura, în urma efectuării măsurătorilor poate obține valorile 95kg și 38,5°C. O a doua categorie de numere, de interes pentru realizarea diferitelor analize, este reprezentată de numărul de pacienți externați dintr-un spital într-o anumită zi, de exemplu 30 de persoane sau numărul de pacienți având vârstele cuprinse în intervalul 65-70 ani, de exemplu 100 de persoane.

La baza redactării acestei cărți se situează două obiective principale: să ofere studenților o modalitate eficientă de organizare, centralizare și prezentare a datelor și, în același timp, să-i sprijine în luarea unor decizii despre o colecție importantă de date prin examinarea unei mici părți reprezentative extrase din aceasta. În vederea atingerii primului obiectiv vor fi utilizate tehnici de analiză statistică descriptivă, iar pentru cel de-al doilea se vor folosi metodele specifice analizei statistice inferențiale.

În primele trei capitole ale cărții sunt prezentate pe larg principalele elemente ale statisticii descriptive, pornind de la detalierea noțiunilor fundamentale ale acestei discipline, modalitățile de prezentare a informațiilor statistice în diferite contexte și conturarea unei prime imagini asupra fenomenului studiat, prin calculul indicatorilor tendinței centrale, ai variației și formei distribuțiilor.

În următoarele trei capitole, au fost introduse primele elemente de statistică inferențială prin trecerea în revistă a noțiunilor de bază ale teoriei probabilităților, completate de prezentarea celor mai des întâlnite distribuții probabilistice discrete (binomială, hipergeometrică și Poisson) și continue (uniformă și normală). De asemenea, tot în această secțiune au fost detaliate caracteristicile de bază ale procesului de eșantionare statistică, prin care se realizează selecția unităților statistice dintr-o populație, în vederea utilizării informațiilor obținute pentru a caracteriza întreaga colectivitate studiată.

A doua parte a cărții, respectiv ultimele șase capitole, realizează o prezentare succintă a principalelor metode specifice analizei statistice inferențiale, începând cu procesul de estimare a parametrilor unei populații, testarea ipotezelor statistice, principalele teste parametrice (testul „z", Student „t", Fisher „F", ANOVA) utilizate în practica medicală, analiza de regresie și corelație simplă liniară precum și testul CHI-pătrat, folosit în mod frecvent pentru evaluarea asocierii dintre două variabile categoriale.

Acest manual își propune să ofere o introducere clară și accesibilă în biostatistică, fiind destinat atât studenților din domeniul științelor vieții, cât și cercetătorilor care doresc să își îmbunătățească abilitățile analitice.

Fie că este vorba despre analiza unui set de date privind eficacitatea unui tratament, despre studiul factorilor de risc ai unei boli sau despre

Scanned with OKEN Scanner

înțelegerea variațiilor biologice din cadrul unei populații, biostatistica oferă un cadru metodologic esențial pentru validarea concluziilor cercetării și evitarea erorilor de interpretare.

Prin exemple practice, explicații intuitive și aplicații relevante, acest manual va ghida cititorul în procesul de învățare a biostatisticii, punând accent pe utilitatea sa practică. Indiferent de pregătirea anterioară și de nivelul de experiență în domeniu, scopul nostru este ca fiecare cititor să își dezvolte competențele necesare pentru a analiza și înțelege datele biostatistice într-un mod critic și temeinic fundamentat.

Exemplele și aplicațiile practice care sunt prezentate în această carte sunt realizate atât prin utilizarea unui program specializat de analiză statistică, așa cum este IBM SPSS Statistics, cât și prin folosirea unui program de calcul tabelar extrem de răspândit și util, respectiv Microsoft Excel.

1.2. DEFINIȚIA ȘI IMPORTANȚA PRACTICĂ A STATISTICII

Inițial o modalitate practică de evidență a populației și bunurilor materiale, atestată de peste 6000 de ani, statistica și-a perfecționat în mod continuu noțiunile, metodele și tehnicile de cercetare, nereprezentând doar un instrument de lucru al altor științe, ci o disciplină științifică de sine stătătoare, care deține un rol esențial în procesul complex de analiză a datelor cât și în luarea unor decizii informate.

Cuvântul "statistică" provine din limba latină, respectiv termenul „status” cu semnificația de "stare" sau "situație". În secolul al XVIII-lea, termenul a fost influențat de cuvântul „Statistik”, introdus de filozoful, istoricul și statisticianul german Gottfried Achenwall (1719-1772), profesor la Universitatea din Göttingen, pentru a desemna știința care studiază informațiile ce contribuie la caracterizarea diferitelor aspecte politice, economice și sociale din viața unei națiuni. În concepția lui Achenwall statistica nu era o știință bazată pe matematică și teoria probabilităților, ci mai degrabă o disciplină descriptivă folosită pentru a analiza structura și organizarea statelor. Cu toate acestea, Gottfried Achenwall a pus bazele statisticii moderne, transformând această disciplină într-un domeniu de studiu academic în cadrul Universității din Göttingen. Ideile sale au fost dezvoltate ulterior de către matematicianul, biologul și statisticianul britanic Karl Pearson (1857–1936), considerat părintele statisticii moderne. Acesta a revoluționat domeniul prin dezvoltarea analizei corelației, conceptului de regresie liniară și testului CHI-pătrat, pentru verificarea asocierii dintre două

Scanned with OKEN Scanner

variabile categoriale, punând în acest fel bazele statisticii inferențiale. Începând din 1884 a fost numit profesor de matematici aplicate și mecanică la University College din Londra, unde a fondat prima catedră de Statistică matematică. În prezent metodele lui Pearson sunt indispensabile în procesul de analiză a datelor, fiind folosite în diferite domenii de cercetare, la realizarea de predicții precum și în dezvoltarea sistemelor de inteligență artificială.

Statistica are un obiect de studiu bine definit, bazat pe cercetarea analitică a datelor, având drept scop extragerea de informații relevante și fundamentarea deciziilor în condiții de incertitudine. În același timp, Statistica utilizează metode proprii de colectare, de organizare și reprezentare a datelor, de analiză statistică descriptivă și inferențială, cu rădăcini profunde în teoria probabilităților dar dezvoltând propriile sale concepte, precum distribuțiile statistice, erorile de eșantionare și testele de semnificație.

Metodele statistice sunt utilizate pe scară largă în multiple domenii, printre care medicină, biologie, economie, sociologie, psihologie, inginerie, inteligență artificială, incluzând o serie de tehnici avansate precum analiza big data și învățarea automată, care au permis statisticii moderne să se dezvolte ca o disciplină științifică autonomă.

Statistica este o știință metodologică care studiază fenomenele și procesele colective din natură și din societate pe baza observării lor științifice, a culegerii și prelucrării detaliate a informațiilor, cu scopul de a formula și explica legile, legitățile și regularitățile care le guvernează.

Ceea ce caracterizează în mod specific obiectul statisticii este faptul că ea studiază fenomenele și procesele de masă din natură și din societate, sub formă de mulțimi sau colectivități, care s-au constituit ca rezultat al acțiunii unui mare număr de cauze cu caracter esențial asociate cu factori întâmplători, ale căror influențe depind de condițiile concrete de spațiu și timp în care s-au realizat.

În situația când datele supuse analizei aparțin domeniilor Biologiei și Medicinei este utilizat termenul de *„biostatistică”*, pentru a face o distincție clară a zonei de aplicabilitate a conceptelor, metodelor și instrumentelor statistice.

Biostatistica este știința care analizează și sintetizează datele statistice medicale și biologice, convertindu-le în informații valoroase, capabile să ghideze deciziile din domeniul sănătății. Fie că vorbim despre înțelegerea mecanismelor bolilor, evaluarea eficienței unui tratament sau anticiparea evoluției unor epidemii, biostatistica stă la baza acestor analize complexe.

Scanned with OKEN Scanner

În acest material didactic, vom explora posibilitatea utilizării metodelor statistice pentru a descifra regulile și tendințele existente în structura profundă a datelor biologice și medicale, oferind suportul necesar unor concluzii temeinic fundamentate, bazate pe seturi de date reale.

Biostatistica este un element fundamental în aplicarea metodei științifice în domeniul biologiei și al sănătății, oferind instrumentele necesare pentru a transforma informațiile și datele brute în cunoștințe utile pentru specialiștii din aceste zone de cercetare.

Metoda științifică este un proces sistematic de investigare a realității utilizat pentru a înțelege fenomenele naturale, a testa diferite ipoteze și a genera cunoștințe noi. Ea presupune o serie de pași structurați, care includ observarea unui fenomen, formularea unei ipoteze, efectuarea experimentelor sau colectarea datelor, analiza rezultatelor și, în final, formularea unor concluzii care pot confirma sau infirma ipoteza inițială de cercetare.

Principalii pași ai metodei științifice sunt:

A. *Observarea*: identificarea unui fenomen sau a grup de fenomene a căror observație generează o serie de întrebări la care se va putea răspunde într-un mod riguros științific (de ex. este ușor de observat faptul că, în cadrul unei diete, persoanele care efectuează regulat și exerciții fizice înregistrează o scădere mai mare în greutate comparativ cu cele care sunt sedentare).

B. *Formularea ipotezei*: în acest al doilea pas, se formulează o ipoteză în scopul explicării fenomenului studiat și pentru a realiza predicții asupra unui nou set de observații. Se pot formula astfel 2 tipuri de ipoteze:
   - Ipoteza de cercetare: „Exercițiile fizice par să reducă greutatea corporală"
   - Ipoteza statistică: Reducerea medie a greutății corporale a persoanelor care efectuează exerciții fizice este mai mare comparativ cu reducerea medie a greutății corporale a persoanelor care nu fac exerciții fizice"

C. *Proiectarea experimentului*: se referă la planificarea și organizarea unui experiment într-o manieră care să asigure obținerea unor date valide și fiabile, permițând testarea corectă a ipotezelor formulate. Designul experimentului implică luarea unor decizii cheie legate de:
   - Definirea variabilelor: stabilirea variabilelor independente (factorii controlați sau manipulați) și a variabilelor dependente (cele care sunt măsurate ca răspuns la modificările variabilelor independente);

Scanned with OKEN Scanner

- Alocarea în mod aleatoriu a participanților sau unităților de studiu în grupuri care primesc tratamentul sau intervenția (grup experimental) și în grupuri care nu primesc tratamentul (grup de control), pentru a compara efectele. Aceasta ne permite, în situația când toți ceilalți factori potențiali sunt controlați (nu influențează rezultatele), să testăm o relație cauzală: în acest fel, vom putea concluziona sau nu că scăderea mai mare a greutății corporale în grupul experimental a fost rezultatul practicării exercițiilor fizice;
- Repetabilitate și validitate: *planificarea experimentului astfel* încât acesta să poată fi replicat de un număr mare de ori până în momentul când va fi credibil din punct de vedere științific.

D. *Analiza datelor*: Utilizarea tehnicilor statistice pentru a evalua rezultatele obținute (ulterior analizei statistice descriptive a variabilei: greutatea corporală a persoanelor din cele 2 grupuri, vom apela la analiza statistică inferențială, respectiv testarea ipotezelor privind diferența mediilor celor două populații, din care provin eșantioanele de lucru, prin utilizarea testului Student, în cazul exemplului de față).

E. *Concluziile, interpretarea rezultatelor și confirmarea sau infirmarea ipotezei de cercetare*. În analiza statistică lucrăm cu două tipuri de ipoteze: H_0 „ipoteza nulă sau ipoteza absenței diferențelor", care va fi testată în scopul expres de a fi discreditată și H_A „ipoteza alternativă", care reprezintă o afirmație „diametral opusă ipotezei nule", pe care o vom considera adevărată doar în situația când datele analizate la nivel de eșantion aduc suficiente dovezi în sprijinul respingerii ipotezei nule H_0. În consecință, varianta diametral opusă a ceea ce dorește să dovedească echipa de cercetare va fi folosită ca enunț al ipotezei nule. Ne putem gândi la ipoteza nulă ca reprezentând teoria clasică (implicită) care necesită dovezi suficient de solide, furnizate de eșantionul supus studiului nostru, pentru a putea fi în consecință respinsă. De exemplu, în situația când comparăm mediile a două populații, ipoteza nulă afirmă faptul că diferența dintre cele 2 medii este egală cu zero, sau, cu alte cuvinte, populațiile supuse analizei nu sunt diferite. Procesul de testare a ipotezelor statistice în particular și chiar inferența statistică în general nu conduc către dovedirea vreunei ipoteze, ci doar indică situația în care o anumită ipoteză statistică este susținută sau nu de datele disponibile la nivel de eșantion.

Scanned with OKEN Scanner

F. *Reproductibilitatea*: Indiferent de concluziile procesului științific, foarte rar rezultatele unui singur studiu sunt considerate a fi concludente. Din acest motiv, este necesară testarea repetată de către alți cercetători, prin studii similare, pentru a confirma rezultatele și a asigura în acest fel respectivului subiect de cercetare o credibilitate mai mare din punct de vedere științific.

1.3. SCURT ISTORIC AL STATISTICII ROMÂNEȘTI

În Țările Române, prima și cea mai relevantă lucrare care a înglobat elemente de statistică a fost monografia cărturarului și domnitorului Moldovei, *Dimitrie Cantemir* (1674 – 1723) – membru al Academiei de Științe din Berlin (1714), intitulată "*Descrierea Moldovei*", redactată în limba latină, în anul 1716, tipărită în limba germană între anii 1769 – 1770, tradusă apoi în românește și publicată în anul 1825 la tipografia Mânăstirii Neamț. Această lucrare a constituit un element de referință în dezvoltarea statisticii pe plan economic, social, politic, geografic și cultural din țara noastră.

Dimitrie Cantemir

Contribuții românești importante la dezvoltarea statisticii descriptive au mai avut:

Nicolae Șuțu (1798 – 1871) cu lucrarea *„Noțiuni statistice despre Moldova"* tipărită în limba franceză la Iași în anul 1849. Cartea este una dintre primele lucrări care oferă o analiză statistică detaliată a Moldovei din secolul al XIX-lea. În cadrul studiului sunt analizate elemente de statistică demografică (estimarea numărului de locuitori, structura populației, distribuția etnică și ocupațiile locuitorilor), economie (producția de cereale, creșterea vitelor precum și rolul economic al moșierilor și țăranilor), finanțe (structura administrativă a Moldovei, veniturile statului și cheltuielile publice, impozitele și dările plătite de locuitori), comerț și infrastructură (principalele rute comerciale, târgurile și piețele Moldovei precum și relațiile comerciale derulate cu Imperiul Otoman, Rusia, etc.). Cartea, bazată pe metode analiză

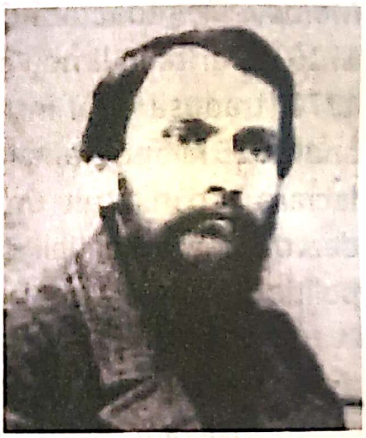
Nicolae Șuțu

Scanned with OKEN Scanner

moderne pentru acele vremuri, a reprezentat prima lucrare cu caracter statistic, fiind un pilon important în procesul de dezvoltare a acestei științe în România.

Ion Ionescu de la Brad

Ion Ionescu de la Brad (1818 – 1891) cu lucrarea *„Povățuiri pentru facerea catagrafiei Moldovei precedate de oare cari elemente de statistică"* publicată în anul 1859, care constituie primul manual de statistică elaborat de un autor român. Aceasta reprezintă prima lucrare metodologică despre principiile și etapele realizării unui recensământ (catagrafie), care a pus accent pe colectarea, analiza și interpretarea riguroasă a datelor, creând astfel bazele organizării statisticii oficiale din țara noastră.



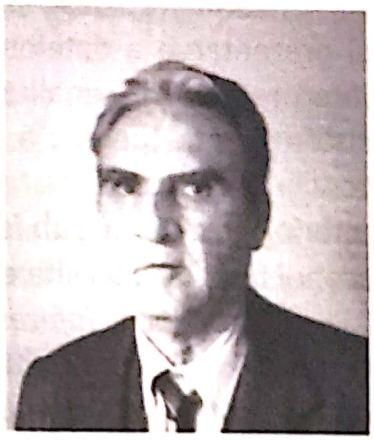
Dionisie Pop Marțian

Dionisie Pop Marțian (1829 – 1865) a organizat și condus primul Oficiu Central de Statistică din Muntenia și după Unire pe cel al Principatelor Unite, fiind în același timp și inițiatorul și conducătorul primei reviste științifice de specialitate din România, apărută în anul 1860 și intitulată *Analele Statistice*. Dionisie Pop Marțian rămâne o figură emblematică în evoluția statisticii românești, fiind un pionier al metodologiei științifice în procesul de analiză a datelor sociale și economice. Marțian a pledat pentru utilizarea metodelor cantitative și a analizelor comparative în statistică, promovând aplicarea matematicii și calculului probabilităților în studiile demografice și economice. De asemenea, a publicat numeroase articole și studii în care a argumentat importanța utilizării statisticii în susținerea procesului de dezvoltare social-economică a României. La Congresul Statistică desfășurat la Berlin în anul 1863, unde a participat ca delegat oficial al țării noastre, Dionisie Pop Marțian a susținut faptul că *„...vom avea în curând o statistică destul de completă pentru ca guvernul să poată întemeia pe datele ei reformele pe care tara le așteaptă cu o vie nerăbdare... căci... statistica este singurul tărâm sigur pe care un guvern prevăzător trebuie să întemeieze rațiunea reformelor sale"*.

Printre personalitățile marcante, care au avut un cuvânt greu de spus în statistica românească, s-au remarcat în special Leonida Colescu (1872-1940), Dimitrie Gusti (1880-1955), Virgil Madgearu (1887-1940), Octav

Scanned with OKEN Scanner

Onicescu (1892-1983), Sabin Manuilă (1894-1964), Anton Golopenția (1909-1951), Nicolae Georgescu-Röegen (1906-1994), Gheorghe Mihoc (1906-1981), Vladimir Trebici (1916-1999), Constantin Ionescu (1926-2013) și Mihai Țarcă (1941-2021).



Mihai Țarcă

Profesor renumit al Universității „Al. I. Cuza" din Iași la disciplinele Statistică și Demografie, șef Catedrei de Statistică și ulterior Decan al Facultății de Științe Economice, *Mihai Țarcă* (1941 – 2011) a contribuit la înființarea în anul 1992 a specializării „Statistică economico-socială" în cadrul Facultății de Științe Economice, care, din anul 1997, a condus la apariția primelor promoții de „statisticieni de profesie" din Moldova.

Formarea sa profesională a fost desăvârșită prin numeroase studii și stagii de documentare efectuate la Humbold Universitat – Berlin, la Georgetown University – Washington DC, Freiburg University – Germania, University of Nebraska at Omaha, University of Texas (Arlington), Universite de Lille – Franța, Reading University – Marea Britanie, University of Nijmegen – Olanda, Universite Tolouse I – Franța.

Profesorul Mihai Țarcă a avut un rol esențial în realizarea și coordonarea mai multor manuale universitare, cărți, monografii și articole științifice valoroase din domeniile statisticii, economiei și demografiei, fiind autorul primului *„Tratat de Demografie"* din România, publicat în Editura Junimea din Iași, în anul 2008.

1.4. TIPURI DE ANALIZĂ STATISTICĂ – DESCRIPTIVĂ ȘI INFERENȚIALĂ

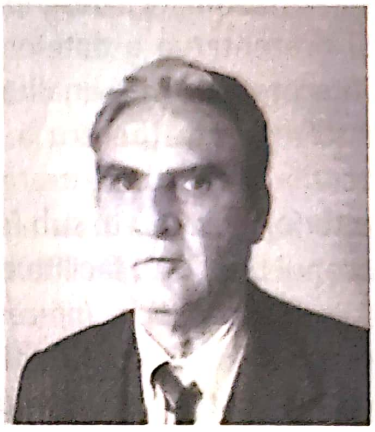
În general, cu ocazia diferitelor studii și cercetări se colectează un volum însemnat de informații privind caracteristicile unităților statistice de interes sau variabilele studiate, care nu pot fi înțelese în profunzime doar prin simpla lor observație.

Informațiile obținute în urma înregistrării fiecărei unități a colectivității cercetate cu caracteristicile ei constituie materialul statistic inițial, "materia primă" a statisticii, pe baza căreia se înfăptuiește cunoașterea științifică a realității.

După obținerea informațiilor statistice și verificarea autenticității lor (prevenirea erorilor de înregistrare și controlul cantitativ și calitativ al datelor

Scanned with OKEN Scanner

Onicescu (1892-1983), Sabin Manuilă (1894-1964), Anton Golopenția (1909-1951), Nicolae Georgescu-Röegen (1906-1994), Gheorghe Mihoc (1906-1981), Vladimir Trebici (1916-1999), Constantin Ionescu (1926-2013) și Mihai Țarcă (1941-2021).



Mihai Țarcă

Profesor renumit al Universității „Al. I. Cuza" din Iași la disciplinele Statistică și Demografie, șef Catedrei de Statistică și ulterior Decan al Facultății de Științe Economice, *Mihai Țarcă* (1941 – 2011) a contribuit la înființarea în anul 1992 a specializării „Statistică economico-socială" în cadrul Facultății de Științe Economice, care, din anul 1997, a condus la apariția primelor promoții de „statisticieni de profesie" din Moldova.

Formarea sa profesională a fost desăvârșită prin numeroase studii și stagii de documentare efectuate la Humbold Universitat – Berlin, la Georgetown University – Washington DC, Freiburg University – Germania, University of Nebraska at Omaha, University of Texas (Arlington), Universite de Lille – Franța, Reading University – Marea Britanie, University of Nijmegen – Olanda, Universite Tolouse I – Franța.

Profesorul Mihai Țarcă a avut un rol esențial în realizarea și coordonarea mai multor manuale universitare, cărți, monografii și articole științifice valoroase din domeniile statisticii, economiei și demografiei, fiind autorul primului „*Tratat de Demografie*" din România, publicat în Editura Junimea din Iași, în anul 2008.

1.4. TIPURI DE ANALIZĂ STATISTICĂ – DESCRIPTIVĂ ȘI INFERENȚIALĂ

În general, cu ocazia diferitelor studii și cercetări se colectează un volum însemnat de informații privind caracteristicile unităților statistice de interes sau variabilele studiate, care nu pot fi înțelese în profunzime doar prin simpla lor observație.

Informațiile obținute în urma înregistrării fiecărei unități a colectivității cercetate cu caracteristicile ei constituie materialul statistic inițial, "materia primă" a statisticii, pe baza căreia se înfăptuiește cunoașterea științifică a realității.

După obținerea informațiilor statistice și verificarea autenticității lor (prevenirea erorilor de înregistrare și controlul cantitativ și calitativ al datelor

Scanned with OKEN Scanner

statistice), gruparea statistică sub forma unor tabele sau reprezentări grafice adecvate facilitează într-o primă etapă conturarea unei imagini mai clare și înțelegerea mai bună a esenței fenomenelor medicale, sociale sau economice studiate.

Statistica descriptivă se referă la metodele de organizare, centralizare și prezentarea a datelor într-o formă cât mai sugestivă, ușor de înțeles și interpretat. Prin analiza statistică descriptivă informațiile obținute sunt sistematizate (pentru a surprinde trăsăturile de bază ale lor, a descoperi tendința de concentrare a acestora, a identifica valorile extreme, etc.) și ulterior prezentate sub forma unor tabele, diagrame și indicatori statistici, în scopul final de a facilita extragerea esenței din setul de date statistice.

Statistica inferențială reprezintă metodele utilizate în vederea estimării unei proprietăți sau caracteristici a unei populații generale, pe baza datelor obținute la nivelul unui eșantion reprezentativ extras din aceasta.

Există numeroase situații când deciziile trebuie luate pe baza unui set limitat de date. De exemplu, în situația în care am dori să cercetăm posibila asociere între statusul fumător / nefumător al mamei în primul trimestru de sarcină (factor de risc) și obezitatea la copilul de 5-6 ani (maladie) în cazul unei anumite regiuni. Această activitate de colectare a informațiilor de la întreaga „populație" a copiilor cu vârstele de 5-6 ani ar necesita un consum extrem de mare de resurse de timp și financiare, fiind aproape imposibil de realizat în practică. Cu toate acestea, utilizând metode ale statisticii inferențiale, putem culege informațiile doar de la un număr limitat de copii din acest interval de vârstă, utilizând un „eșantion reprezentativ " extras din populația generală (figura 1.1).

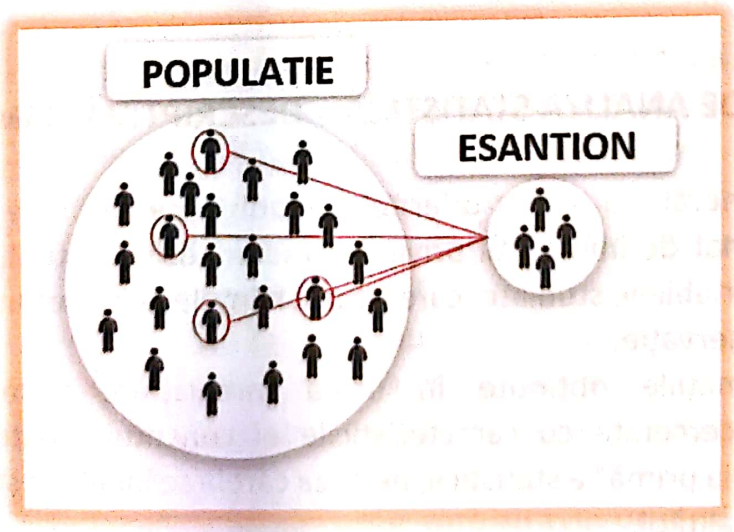


Figura 1.1. Populația generală și eșantionul

Scanned with OKEN Scanner

Populația reprezintă întregul set de indivizi, obiecte sau fenomene de interes împreună cu caracteristicile lor care fac obiectul unei cercetări statistice. Un indicator statistic de sinteză calculat la nivelul unei populații se numește *parametru*. De exemplu, dacă ne referim la variabila statistică „vârstă", parametrul este reprezentat de vârsta medie a întregii populații care prezintă interes în cadrul cercetării.

Eșantionul reprezintă doar o mică parte a populației generale studiate, care reproduce într-un mod cât mai fidel trăsăturile acesteia (o reproducere miniaturală a populației). Un indicator statistic de sinteză calculat la nivelul unui eșantion poartă denumirea de *statistică*. Mergând pe linia exemplului anterior, statistica va fi reprezentată de vârsta medie a eșantionului reprezentativ extras din populația generală, care va fi utilizată ulterior în estimarea parametrului corespunzător de la nivel populațional, respectiv vârsta medie a întregii populații.

Elementele componente ale colectivităților (populații sau eșantioane) care se înregistrează cu trăsăturile, însușirile sau proprietățile lor principale poartă denumirea de unități statistice.

Unitățile statistice sunt ființe, lucruri, fapte sau evenimente elementare, care fac obiectul observației statistice, indiferent de natura lor, cu condiția de a răspunde cerințelor unei definiții precise. Totalitatea unităților statistice care fac parte din cadrul unei colectivități (populație sau eșantion) formează *volumul* acesteia.

1.5. CATEGORII DE VARIABILE UTILIZATE ÎN ANALIZA STATISTICĂ

Variabila statistică este acea caracteristică sau proprietate a unui individ, obiect sau fenomen, de interes pentru o cercetare statistică, care este măsurabilă și poate lua diferite valori de la o unitate statistică la alta. Astfel, în momentul când observăm o caracteristică care poate lua diverse valori în cazul diferitelor persoane, obiecte sau fenomene cercetate vom utiliza denumirea de *variabilă*. De exemplu, în domeniul medical putem întâlni o varietate de variabile statistice care pot prezenta interes în cadrul diferitelor studii, cum ar fi: tensiunea arterială, pulsul, glicemia, colesterolul, grupa de sânge, VSH-ul, scorul APGAR, IMC-ul, albumina, hemoglobina, genul, greutatea, înălțimea, vârsta, mediul social, etc.

Variabila statistică reprezintă însușirea, trăsătura sau proprietatea principală, comună tuturor unităților statistice ale unei colectivități, ale cărei valori diferă, în general, de la o unitate statistică la alta sau de la un grup de unități la altul.

Scanned with OKEN Scanner

După modul de exprimare, caracteristicile pot fi:

- *numerice* (cantitative) – măsurabile și exprimabile numeric, cum ar fi: pulsul, glicemia, colesterolul, IMC-ul, albumina, hemoglobina, greutatea, înălțimea, vârsta, etc.;
- *atributive* (calitative) – se exprimă prin cuvinte: genul, grupa de sânge, mediul social, starea civilă, culoarea ochilor, etc.

Diversele stări pe care le pot poseda caracteristicile calitative și diferitele valori numerice pe care le pot lua caracteristicile cantitative se numesc variante sau modalități ale lor.

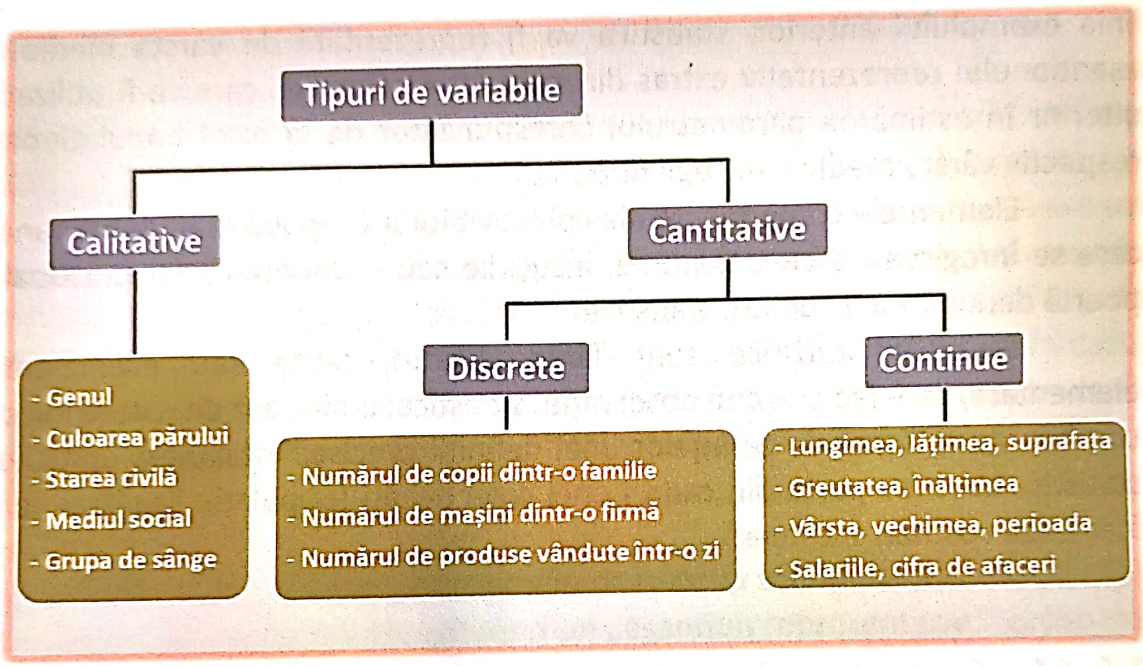


Figura 1.2. Tipuri de variabile statistice

Sintetizând, principalele tipuri de variabile statistice pe care le întâlnim în practică, în funcție de modalitatea concretă de exprimare, sunt prezentate în figura 1.2.

În cazul variabilelor calitative (categoriale), de exemplu starea civilă, vom determina ponderile (%) deținute de populația căsătorită, necăsătorită, divorțată și văduvă.

Variabilele cantitative sunt cele care se pot raporta numeric și pot fi *discrete* (iau doar anumite valori, între care există spații goale în care nu există nicio altă valoare posibilă) sau *continue* (pot lua absolut orice valoare în cadrul unui anumit interval). De exemplu, numărul pacienților internați zilnic într-un spital reprezintă o variabilă discretă, deoarece valorile pe care aceasta le poate înregistra sunt reprezentate de numere întregi: 1, 2, 3, 4 sau 5, fiind

Scanned with OKEN Scanner

imposibil de obținut valori cum ar fi 1,70 sau 3,25. În situația variabilelor continue însă, cum ar fi greutatea pacienților, indiferent cât de mult se apropie valorile pentru 2 persoane diferite, vom putea, cel puțin din punct de vedere teoretic, să identificăm un pacient a cărui greutate să se interpună între cele două valori. Trebuie precizat însă faptul că, datorită limitelor sau imperfecțiunilor evidente ale instrumentelor de măsură observațiile realizate asupra variabilelor continue sunt în final înregistrate într-o formă mai mult sau mai puțin discretă.

În funcție de rolul lor deținut în planificarea și organizarea unui experiment științific, variabilele statistice pot fi:

- *independente* (factorii controlați sau manipulați de cercetător în cadrul experimentului): de exemplu tipul de medicament administrat în cadrul unui tratament (medicament nou vs. placebo);
- *dependente* (variabilele care sunt măsurate ca răspuns la modificările variabilelor independente): de exemplu modificarea valorilor tensiunii arteriale sistolice și diastolice după 8 săptămâni de tratament;
- *de control* (factori care ar putea influența rezultatele unui experiment și care trebuie menținuți constanți pentru a asigura validitatea studiului): de exemplu: vârsta, sexul, IMC-ul, alimentația, activitatea fizică, alte medicamente, consumul de alcool si tutun, nivelul de stres și somnul, care ar putea interfera cu sistemul de tratament, putând denatura rezultatele experimentului.

În situația în care variantele unei anumite variabile statistice apar datorită „șansei", acestea neputând fi prezise în avans, respectiva variabilă poartă denumirea de *variabilă aleatorie*. Cu titlu de exemplu, înălțimea unui adult este o variabilă aleatorie, întrucât în momentul nașterii copilului nu este posibilă prezicerea înălțimii exacte a viitorului adult, la maturitate, aceasta fiind rezultatul acțiunii complexe a unui număr însemnat de factori genetici și de mediu.

1.6. NIVELURILE DE MĂSURARE A DATELOR

Determină modalitatea în care rezultatele diverselor măsurători efectuate asupra caracteristicilor unităților statistice din cadrul colectivităților studiate pot fi organizate, centralizate și prezentate în funcție de natura lor și

Scanned with OKEN Scanner

în același timp și tipul de analiză statistică care va fi utilizat în prelucrarea lor ulterioară.

Procesul de măsurare în esență reprezintă atribuirea unor valori numerice caracteristicilor persoanelor, obiectelor sau evenimentelor care prezintă interes pentru o cercetare statistică, în baza unor seturi de reguli bine definite.

În practică deosebim patru niveluri de măsurare a datelor statistice, pornind de la cel de bază, în cazul datelor de tip nominal, până la cel mai complex, în situația datelor de tip raport (figura 1.3).

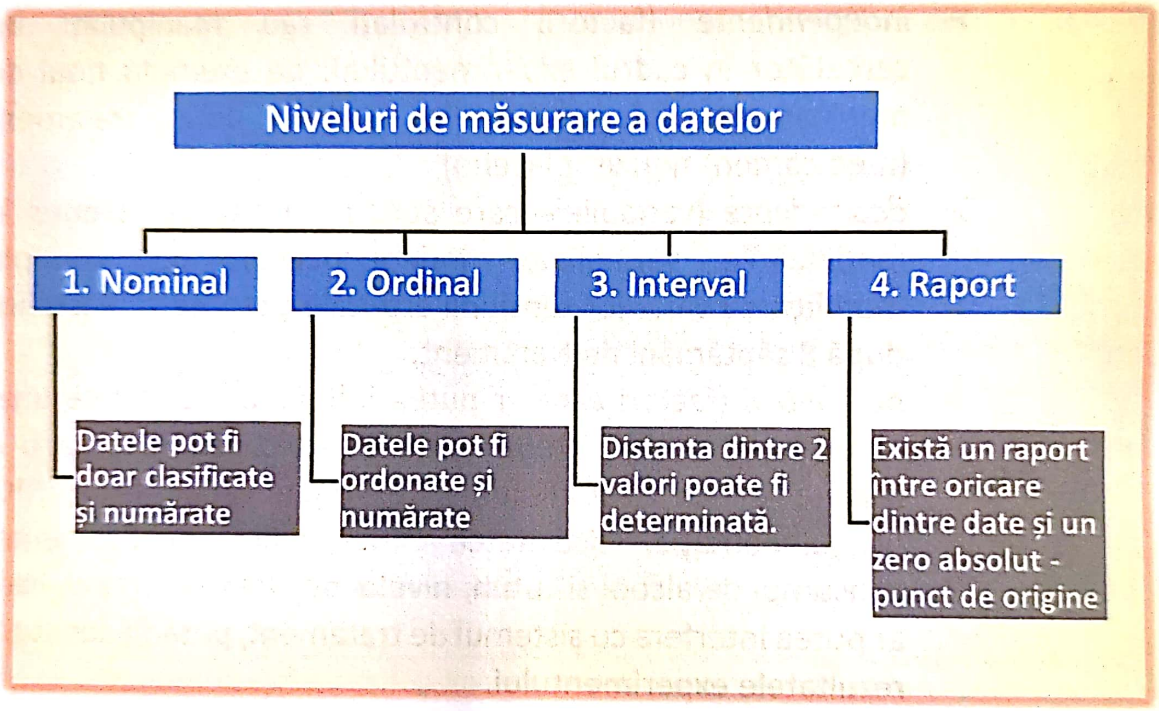

Niveluri de măsurare a datelor
1. Nominal
2. Ordinal
3. Interval
4. Raport
Datele pot fi doar clasificate și numărate
Datele pot fi ordonate și numărate
Distanta dintre 2 valori poate fi determinată.
Există un raport între oricare dintre date și un zero absolut - punct de origine

Figura 1.3. Nivelurile de măsurare a datelor statistice

Putem exemplifica nivelul nominal al datelor statistice prin distribuția populației rezidente a României după variabila „stare civilă" la recensământul populației și locuințelor din 2021 (tabelul 1.1). Vom realiza o clasificare a nivelurilor variabilei „stare civilă", fără ca între acestea să putem defini o ordine evidentă (ordonarea celor patru categorii este nerelevantă în acest caz).

Scanned with OKEN Scanner

Tabelul 1.1. Distribuția populației României după starea civilă la recensământul populației și locuințelor din 2021

Starea civilă	Persoane	% din total
Necăsătorit/ă	7.282.510	38,2%
Căsătorit/ă	8.591.549	45,1%
Văduv/ă	1.803.029	9,5%
Divorțat/ă	1.376.727	7,2%
Total	**19.053.815**	**100,0%**

Sursa: prelucrat după https://www.recensamantromania.ro/rezultate-rpl-2021/rezultate-definitive-caracteristici-demografice/

Datele de tip ordinal sunt specifice tot variabilelor statistice calitative, cu precizarea suplimentară că acestea se pot împărți în diferite categorii, care însă pot fi ordonate pe o scală ordinală (de exemplu: insuficient, suficient, mediu, bine, foarte bine). O caracteristică importantă a datelor de tip ordinal se referă la imposibilitatea comparării magnitudinii diferențelor dintre categoriile adiacente.

De exemplu, utilizăm distribuția cu date fictive a calificativelor celor 100 de studenți din anul I la disciplina Biostatistică (tabelul 1.2). Datele sunt centralizate în 5 categorii, ordonate în funcție de calificativul obținut pentru prezentarea proiectului. De asemenea, putem converti valorile absolute ale frecvențelor fiecărei clase/categorii în procente. Categoriile reflectă calitatea nivelului de pregătire a studenților, însă nu putem afirma că diferența între studenții care au obținut „foarte bine" și „bine" ar fi la fel comparativ cu cea între cei care au obținut „mediu" și „slab".

Tabelul 1.2. Distribuția studenților din anul I la examen

Calificativ	Nr. studenți	% din total
Foarte bine	8	8,0%
Bine	26	26,0%
Mediu	42	42,0%
Slab	19	19,0%
Foarte slab	5	5,0%
Total	**100**	**100.0%**

Scanned with OKEN Scanner

Datele de tip interval reprezintă al treilea nivel ierarhic de măsurare, care include toate caracteristicile nivelului anterior (ordinal), iar în plus intervalul sau diferența dintre oricare 2 valori alăturate este aceeași. De asemenea, prezența lui zero este arbitrară, reprezentând doar un punct pe scară, care nu semnifică absența condiției. Ne putem gândi la scara de temperatură Celsius, în care diferitele niveluri ale temperaturii pot fi ordonate crescător sau descrescător și putem calcula orice interval între două niveluri ale temperaturii (1 grad Celsius reprezintă o unitate de măsură constantă). Astfel distanța dintre 5 și 10 grade Celsius este egală cu cea dintre 40 și 45 de grade, respectiv 5 grade Celsius. Valoarea zero reprezintă doar un punct pe scara de temperatură, care nu arată absența oricărui nivel al acesteia (căldură sau frig).

Datele de tip raport reprezintă cel mai complex nivel de măsurare a variabilelor statistice , având toate proprietățile datelor de tip interval și în plus un „zero absolut", tratat ca punct de origine. Banii sunt un exemplu foarte potrivit. În situația când o persoană are zero lei, înseamnă că nu are niciun ban asupra sa și în același timp, atunci când un salariat câștigă 10000 lei / lună, acest nivel reprezintă dublu comparativ cu alt salariat care câștigă doar 5000 lei / lună. De asemenea, greutatea sau înălțimea sunt măsurate pe o scară de tip raport, nivelul zero exprimând absența oricărei valori a greutății sau înălțimii. În același timp, o persoană de 150 kg cântărește triplu față de una de 50kg.

Sintetizând, principalele caracteristici ale nivelurilor de măsurare a unei variabile statistice sunt evidențiate în figura 1.4.

Nivelul de măsurare a datelor va determina în mod concret metodele statistice care pot fi utilizate în practică în vederea analizei diferitelor variabile.

Nivelul de măsurare a datelor statistice	Nominal	Ordinal	Interval	Raport
Este stabilită ordinea variantelor variabilei?	-	DA	DA	DA
Diferența dintre variante poate fi evaluată?	-	-	DA	DA
Adunarea și scăderea variantelor variabilei?	-	-	DA	DA
Înmulțirea și împărțirea variantelor variabilei?	-	-	-	DA
Există un zero absolut (punct de origine)?	-	-	-	DA

Figura 1.4. Caracteristicile nivelurilor de măsurare a datelor

Scanned with OKEN Scanner

De exemplu, în vederea testării asocierii între două variabile calitative (de nivel nominal și/sau ordinal) va fi utilizat testul neparametric HI pătrat. Într-un alt context, pentru a testa ipoteza statistică privind diferența mediilor a două populații în cazul variabilelor cantitative (de tip interval/raport), vom utiliza în practică testul parametric Student.

1.7. VARIABILITATEA DATELOR MEDICALE

Cercetarea biologică se bazează pe rezultate obținute pe un număr limitat de observații (eșantion) din multiplele posibile (populația), este deci o cercetare prin sondaj. Scopul final al acestei cercetări este reprezentat de generalizarea observațiilor realizate pe un număr limitat de cazuri, la întreaga colectivitate studiată, obținând astfel legi cu aplicabilitate generală.

Acest proces de generalizare depinde de două caracteristici ale datelor statistice:

- *numărul observațiilor studiate* (volumul eșantionului sau lotului analizat);
- *împrăștierea sau dispersia* acestora (varianța).

Datorită specificului lor, datele biologice prezintă o variație proprie, care implică o analiză particulară axată pe teoria probabilităților. În procesul de cercetare și investigare a realității însă apar și diferențe între valorile reale și cele măsurate care au la bază erorile generate atât de metoda metrologică utilizată cât și de precizia limitată a personalului uman implicat.

În cadrul procesului de cercetare științifică din domeniul medical, obiectivul principal este reprezentat de studierea variației biologice, specifice datelor medicale, pe fondul minimizării erorilor care apar inevitabil în derularea oricărui studiu. Variabilitatea datelor din domeniul medical care are la bază erorile de măsurare a variabilelor statistice este caracterizată sintetic prin exactitatea și precizia înregistrării.

Inexactitatea se datorează capacității limitate de măsurare a cercetătorului și depinde de diferiți factori perturbatori, precum modificările temperaturii, cele ale câmpului electric sau magnetic, etc.

Exactitatea reprezintă apropierea valorii numerice determinate experimental de valoarea adevărată (este de fapt eroarea absolută).

Precizia se referă la gradul de fidelitate specific al unei măsurători, fiind dependentă de gradul de sensibilitate a aparatului de măsură utilizat în procesul de colectare a informațiilor. O anumită metodă este considerată a fi precisă în situația când rezultatele determinărilor sunt reproductibile, sau apropiate ca valoare în contextul unor experimente repetate. Precizia este

Scanned with OKEN Scanner

caracterizată de numărul zecimalelor obținute corect în cadrul unei măsurători efectuate.

1.8. ERORILE DE ÎNREGISTRARE

În urma înregistrării informațiilor statistice cu privire la fenomenele colective pot să apară diferențe între nivelurile înregistrate ale unor anumite caracteristici și nivelurile lor reale, diferențe care poartă denumirea de erori de înregistrare.

Cauzele producerii erorilor de înregistrare sunt numeroase:

- lipsa de experiență a personalului care înregistrează;
- ținerea inexactă a evidențelor care stau la baza înregistrărilor;
- erorile de măsurare, etc.

După natura lor, erorile de înregistrare pot fi:

- erori de înregistrare *premeditate* (intenționate);
- erori de înregistrare *nepremeditate* (neintenționate), care la rândul lor pot fi:
  - *erori întâmplătoare*, care nu influențează esențial rezultatele medii, deoarece acțiunile lor într-un anumit sens sunt egal probabile cu acțiunile lor în sens contrar și se compensează reciproc la nivelul colectivității;
  - *erori sistematice*, care acționează într-o anumită direcție, bine determinată și denaturează întotdeauna realitatea.

Controlul datelor statistice este o acțiune ulterioară înregistrării și se efectuează cu scopul de a asigura volumul complet și calitatea datelor înregistrate.

Controlul datelor statistice poate fi:

- control cantitativ (se verifică dacă este complet volumul informațiilor statistice, în sensul înregistrării tuturor unităților statistice, împreună cu caracteristicile lor care prezintă interes în cadrul studiului);
- control calitativ (se verifică înregistrarea corectă a indicatorilor stabiliți: de exemplu, în cazul variabilelor cantitative se verifică existența unor valori neobișnuite, extreme, care depășesc plaja standard de variație; în situația variabilelor calitative, se verifică modul de înregistrare a valorilor în categoriile standard, precum și codificarea corectă a tuturor datelor colectate).

La rândul lui controlul calitativ se poate împărți în:

Scanned with OKEN Scanner

- control logic (se verifică dacă legătura dintre diferite caracteristici înregistrate corespunde unei legături reale);
- control aritmetic (se verifică calculele din care au rezultat totalurile și sub totalurile înregistrate).

Scanned with OKEN Scanner

Capitolul. 2

PREZENTAREA INFORMAȚIILOR STATISTICE

SUBIECTE:

2.1. Procesul de centralizare și grupare a informațiilor statistice
2.2. Construirea distribuțiilor de frecvențe și a graficelor statistice în cazul variabilelor cantitative - histograma și poligonul frecvențelor
2.3. Tabelele de frecvențe și reprezentarea grafică a variabilelor calitative - diagramele prin bare și graficele de structură
2.4. Prezentarea relației sau asocierii dintre două variabile cantitative - diagrama tip „nor de puncte"
2.5. Prezentarea relației sau asocierii dintre două variabile calitative - tabelele de contingență și graficul prin coloane
2.6. Prezentarea relației sau asocierii dintre o variabilă cantitativă și una calitativă - piramida vârstelor

La finalizarea acestui capitol, studentul va fi capabil să:
O2-1: centralizeze variabilele calitative prin realizarea tabelelor de frecvențe (absolute și relative)
O2-2: să reprezinte grafic tabelele de frecvențe (graficele de structură sau diagramele prin bare)
O2-3: centralizeze variabilele cantitative prin construirea distribuțiilor de frecvențe (absolute și relative)
O2-4: reprezinte grafic o distribuție de frecvențe (histograma și poligonul frecvențelor)
O2-5: să creeze și să explice diagramele tip nor de puncte și tabelele de contingență

2.1. PROCESUL DE CENTRALIZARE ȘI GRUPARE A INFORMAȚIILOR STATISTICE

În acest capitol vor fi introduse o serie de proceduri de bază utilizate în analiza descriptivă a datelor statistice. În general, cu ocazia diferitelor studii și cercetări se colectează un volum însemnat de informații privind caracteristicile unităților statistice de interes, care nu pot fi înțelese în profunzime doar prin simpla lor observație.

Informațiile obținute în urma înregistrării fiecărei unități a colectivității cercetate cu caracteristicile ei constituie materialul statistic inițial, "materia

Scanned with OKEN Scanner

primă" a statisticii, pe baza căreia se înfăptuiește cunoașterea științifică a realității (tabelul 2.1).

Tabelul 2.1. Tabelul cu datele a 100 de pacienți, prezentând diferite tipuri de variabile statistice de interes în cadrul cercetării

Nr. crt.	Vârsta (ani)	Gen	Grupa de sânge	Glicemie (mg/dl)	Colesterol (mg/dl)	HTA	Diabet
1	62	Feminin	A	159	180	NU	DA
2	62	Feminin	A	85	145	NU	NU
3	54	Feminin	B	154	238	DA	DA
4	38	Feminin	0	153	200	DA	DA
5	57	Masculin	0	156	257	DA	DA
6	40	Feminin	0	136	170	NU	DA
7	23	Masculin	0	154	195	NU	DA
8	48	Masculin	AB	101	182	NU	NU
9	59	Masculin	A	105	184	DA	NU
10	37	Masculin	B	80	172	NU	NU
...	...	...	...	...	...	...	...
98	41	Feminin	B	96	169	DA	NU
99	56	Masculin	B	117	221	NU	DA
100	52	Masculin	B	94	177	DA	NU

După obținerea informațiilor statistice și verificarea autenticității lor (prevenirea erorilor de înregistrare și controlul cantitativ și calitativ al datelor statistice), gruparea statistică sub forma unor tabele sau realizarea unor grafice adecvate facilitează într-o primă etapă conturarea unei imagini mai clare și înțelegerea mai bună a esenței fenomenelor medicale studiate.

Gruparea statistică reprezintă împărțirea colectivităților statistice (populații sau eșantioane) în grupe omogene de unități, după variația uneia sau a mai multor caracteristici esențiale, comune tuturor unităților colectivității (se mai numește stratificare sau clasificare).

Caracteristica după a cărei variație se repartizează unitățile colectivității în grupe se numește *caracteristică de grupare* (cea care exprimă cel mai clar esența fenomenului studiat și care răspunde cel mai bine scopului cercetării).

Reguli elementare specifice procesului de grupare:

- gruparea unităților statistice (respectiv a variantelor unei variabile) este rațională doar dacă numărul variantelor variabilei de grupare este suficient de mare;

Scanned with OKEN Scanner

- grupările nu prezintă interes decât dacă numărul de grupe sau clase este relativ redus (de obicei până la 15);
- numărul de grupe trebuie să fie suficient de mare pentru a scoate în evidență particularitățile distribuției unităților statistice în cadrul colectivității studiate.

2.2. CONSTRUIREA DISTRIBUȚIILOR DE FRECVENȚE ȘI A GRAFICELOR STATISTICE ÎN CAZUL VARIABILELOR CANTITATIVE - HISTOGRAMA ȘI POLIGONUL FRECVENȚELOR

Un prim pas în derularea fenomenului de organizare și centralizare a datelor statistice cantitative este reprezentat de ordonarea acestora, prin realizarea unei liste a tuturor valorilor variabilei în ordinea mărimii, de la valoarea minimă către cea maximă. Această matrice ordonată permite examinarea rapidă a valorilor minime și maxime ale variabilei precum și concentrarea anumitor valori în anumite zone ale acesteia.

În tabelul 2.2, variabila „Vârstă" este prezentată într-o formă ordonată de la nivelul minim către cel maxim al acesteia.

Tabelul 2.2. Variabila „Vârstă" din tabelul 1 ordonată crescător

17	18	22	22	23	23	25	25	26	27
28	28	29	31	31	31	32	32	32	33
34	35	35	36	36	37	37	38	38	39
39	40	40	40	41	41	41	41	42	42
43	44	44	45	46	46	46	47	47	48
48	48	48	49	49	49	49	49	50	51
51	51	52	52	53	53	54	54	54	54
55	55	55	55	56	57	58	58	58	58
59	59	59	60	60	61	62	62	63	64
64	64	68	68	69	70	70	73	74	75

Prin analiza sumară a informațiilor din tabelul 2.2, observăm imediat vârsta celui mai tânăr pacient (17 ani) și a celui mai vârstnic (75 ani). De asemenea, remarcăm faptul că jumătate dintre pacienți au vârsta sub 48 de ani. Deși aceste observații sunt foarte utile în cadrul procesului de analiză, sunt necesare și alte tehnici statistice, specifice care să conducă la o înțelegere aprofundată a fenomenului studiat. Astfel, pentru a centraliza pacienții din

Scanned with OKEN Scanner

eșantionul studiat în funcție de vârsta fiecăruia, avem nevoie să construim o distribuție a frecvențelor de apariție a fiecărei variante a variabilei statistice „Vârstă".

Așa cum am prezentat anterior, *gruparea statistică* implică împărțirea colectivităților statistice în grupe sau clase omogene de unități, în funcție de variația uneia sau a mai multor variabile statistice de interes în cadrul cercetării (înălțime, greutate, vârstă, gen, grupa de sânge, etc.).

Distribuția de frecvențe reprezintă gruparea tuturor variantelor unei variabile statistice cantitative în intervale distincte, evidențiind numărul de observații sau variante din fiecare clasă sau interval (tabelul 2.3). Intervalele trebuie să fie disjuncte și consecutive, astfel încât de la valoarea minimă până la cea maximă să fie acoperită întreaga plajă de valori a variabilei studiate.

Numărul de indivizi care aparțin unui interval (clasă) formează *frecvența absolută* a intervalului $n_1, n_2, n_3, \dots, n_k$, unde $\sum_{i=1}^{k} n_i = n$ reprezintă volumul lotului analizat.

Tabelul 2.3. Distribuția frecvențelor unei variabile cantitative având caracteristica repartizată pe intervale de grupare

Variabila *X* $(x_{i-1} - x_i)$	**Frecvența absolută** (n_i)	**Frecvența relativă** (f_i)	**Frecvența absolută cumulată** $(N_i \downarrow)$	**Frecvența relativă cumulată** $(F_i \downarrow)$
$x_0 - x_1$	n_1	$f_1 = n_1/n$	$N_1 = n_1$	$F_1 = f_1$
$x_1 - x_2$	n_2	$f_2 = n_2/n$	$N_2 = n_1 + n_2$	$F_2 = f_1 + f_2$
$x_2 - x_3$	n_3	$f_3 = n_3/n$	$N_3 = n_1 + n_2 + n_3$	$F_3 = f_1 + f_2 + f_3$
⋮	⋮	⋮	⋮	⋮
$x_{i-1} - x_i$	n_i	$f_i = n_i/n$	$N_i = n_1 + \dots + n_i$	$F_i = f_1 + \dots + f_i$
⋮	⋮	⋮	⋮	⋮
$x_{k-1} - x_k$	n_k	$f_k = n_k/n$	$N_K = n$	$F_k = 1$
Total	$\sum_{i=1}^{k} n_i = n$	$\sum_{i=1}^{k} f_i = 1$	-	-

Frecvența relativă a unui interval de grupare se calculează prin raportarea frecvenței absolute n_i la numărul de indivizi $f_i = {}^{n_i}/_{n}$.

Frecvența cumulată reprezintă numărul de indivizi cumulați până la un anumit nivel al variabilei de interes, aceasta putând fi la rândul ei absolută și relativă.

Scanned with OKEN Scanner

Numărul de intervale în care se împarte variabila continuă studiată se poate determina în mod aproximativ cu formula lui *Sturges*:

$$k = 1 + log_2\, n \tag{2.1}$$

unde n este volumul eșantionului (în practică utilizează regula 2^k, unde k este numărul de intervale).

Dacă se cunoaște numărul de intervale de grupare k, pentru o colectivitate (populație sau eșantion), mărimea acestora d se poate determina astfel:

$$d = (x_{max} - x_{min})/k \tag{2.2}$$

Intervalele de grupare au de obicei dimensiuni egale, cu excepția celor extreme care au o dimensiune mai mare, pentru a acoperi întreaga plajă de valori.

Revenind la exemplul cu variabila statistică „Vârstă" vom realiza gruparea și centralizarea tuturor celor 100 de variante ale acesteia prin construirea unei distribuții de frecvențe (tabelul 2.4), parcurgând următoarele etape:

A) Stabilirea numărului de intervale de grupare (k): este necesar să determinăm numărul de clase de grupare, utilizând *regula* 2^k → trebuie identificată cea mai mică valoare a lui k pentru care $2^k > n$ (numărul de observații). În cazul exemplului nostru $2^7 = 128 > 100$. Deoarece $2^6 = 64 < 100$, vom lua în considerare valoarea $k = 7$ intervale de grupare.

B) Determinarea mărimii intervalului de grupare (d), utilizând formula 2:

$$d = (x_{max} - x_{min})/k = (75 - 17)/7 = 8{,}29 \text{ ani}$$

În practică, valoarea obținută se rotunjește convenabil (multiplu de 5, 10 sau 100), astfel încât vom considera mărimea intervalului de 10 ani.

C) Precizarea exactă a limitelor fiecărui interval pentru ca fiecare observație să aparțină unei singure clase, evitând astfel suprapunerile.

Tabelul 2.4. Distribuția frecvențelor absolute și relative în cazul variabilei „Vârstă", pe intervale de grupare de 10 ani

Vârsta	Frecvența absolută	Frecvența relativă
10 - 20	2	0,02
20 - 30	11	0,11
30 - 40	18	0,18
40 - 50	27	0,27
50 - 60	25	0,25
60 - 70	12	0,12
70 - 80	5	0,05
TOTAL	**100**	**1,00**

Scanned with OKEN Scanner

În urma centralizării datelor privind variabila cantitativă „Vârsta", observăm faptul că vârstele celor 100 de pacienți, cuprinse între 17 și 65 de ani sunt concentrate în zona mediană: 30 - 60 ani (aproximativ 70% dintre pacienți).

Pentru fiecare interval, putem determina o vârstă tipică situată la mijlocul fiecăruia, respectiv la jumătatea distanței dintre limitele superioară și inferioară. Punctul central al fiecărei clase (interval) aproximează vârstele tuturor pacienților din acel interval, fiind astfel o valoare de referință a acestora.

Vârsta celor mai mulți pacienți (0,27 sau 27% din total) s-a situat in intervalul 40 – 50 ani, al cărui mijloc este valoarea de 45 ani. Putem astfel afirma, faptul că vârsta tipică în intervalul cu frecvență maximă a fost de 45 ani.

În funcție de obiectivele cercetării, poate fi de interes transformarea frecvențelor absolute aferente fiecărui interval de grupare în frecvențe relative, pentru a evidenția proporția din totalul numărului de observații deținută de fiecare clasă (tabelul 2.4). Aceasta se poate realiza prin împărțirea frecvențelor absolute ale fiecărui interval, determinate anterior, la numărul total de observații (*distribuția frecvențelor relative*).

Trebuie ținut cont și de faptul că, transformarea informațiilor inițiale despre vârste într-o distribuție de frecvențe, cu caracteristica grupată pe intervale, are ca rezultat pierderea parțială a unor informații foarte detaliate, sistematizarea datelor nemaiputând evidenția vârsta exactă a fiecărui pacient analizat.

Reprezentarea grafică a informațiilor statistice este o metodă foarte utilă folosită în practică pentru a ușura analiza și generalizarea rezultatelor, aceasta punând în valoare mai bine informațiile sistematizate inițial sub forma distribuțiilor de frecvențe. Elaborarea grafice relevante are ca scop final evidențierea mai clară a diferitelor părți ale colectivității studiate, pe baza variației variabilei și a frecvențelor corespunzătoare. Cele trei tipuri principale de reprezentări grafice care se folosesc în cazul distribuțiilor de frecvențe sunt: histograma, poligonul frecvențelor și poligonul frecvențelor cumulate.

Histograma reprezintă un grafic în care intervalele în care este grupată variabila cantitativă sunt poziționate pe axa orizontală și frecvențele aferente fiecărui interval sunt prezentate pe axa verticală sub forma unor benzi având înălțimea proporțională cu frecvența intervalului corespunzător (figura 2.1). Intervalele vor fi consecutive și nu se vor suprapune. Suprafața totală a graficului va fi egală cu 1 în cazul utilizării frecvențelor relative și cu volumul eșantionului pentru cele absolute.

Histograma pentru o distribuție de frecvențe este similară cu diagrama prin bare, care va fi utilizată în cazul prezentării distribuției variabilelor calitative.

Scanned with OKEN Scanner

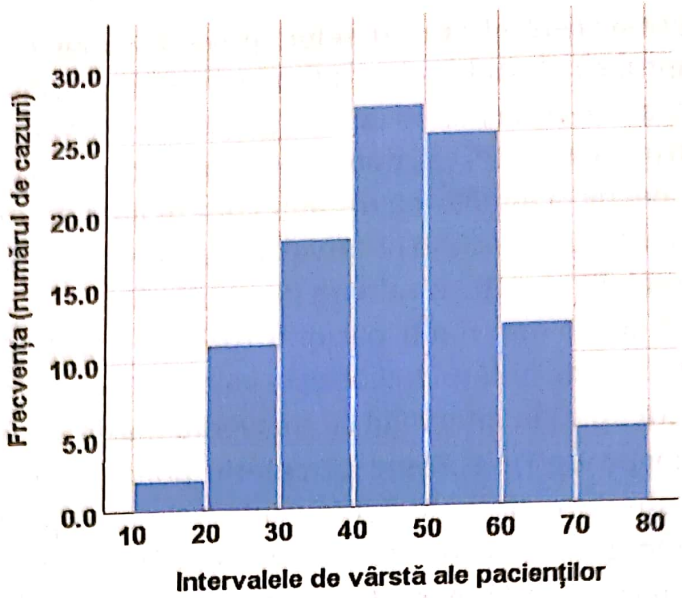


Figura 2.1. Histograma vârstelor celor 100 de pacienți prezentați în tabelul 2.4, pe intervale de grupare de 10 ani

Intervalele de vârstă ale pacienților se situează pe axa absciselor, având aceeași dimensiune (10 ani). În cadrul suprafeței delimitate de histogramă și axa orizontală, fiecare bandă verticală conține o porțiune proporțională cu frecvența respectivului interval. De exemplu, banda corespunzătoare celui de-al doilea interval (20-30 ani) deține $11/100 = 0{,}11$ sau 11% din suprafața totală a histogramei, valoare care coincide cu frecvența relativă a acestuia.

Poligonul frecvențelor este un grafic prin linii care prezintă forma unei distribuții de frecvențe, fiind similar cu histograma (figura 2.2).

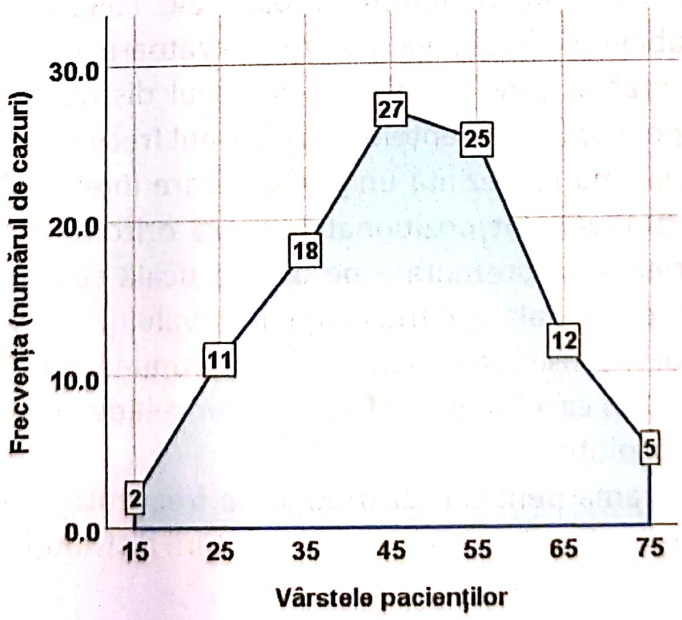


Figura 2.2. Poligonul frecvențelor vârstelor celor 100 de pacienți prezentați în tabelul 2.4

Scanned with OKEN Scanner

Poligonul frecvențelor se obține prin unirea punctelor formate la intersecția mijlocului fiecărui interval cu frecvența corespunzătoare a acestuia. Mijlocul fiecărui interval se obține, ca medie aritmetică simplă a limitelor lui, considerând faptul că variabila studiată este distribuită liniar-uniform în interior. De exemplu, prima valoare a vârstei pacienților, respectiv 15 ani, reprezintă mijlocul intervalului (10 – 20) și se obține ca medie aritmetică a celor două limite 10 și 20, astfel: $15 = (10 + 20)/2$.

Atât histograma cât și poligonul frecvențelor ne permit să ne facem o imagine rapidă a caracteristicilor principale ale datelor (valori minime, maxime, puncte de concentrare a valorilor variabilei), fiecare dintre acestea având o serie de avantaje specifice.

Deși cele două tipuri de reprezentări grafice servesc unui scop similar, histograma prezintă avantajul evidențierii fiecărei clase (interval) sub forma unei benzi verticale, având înălțimea proporțională cu numărul valorilor variabilei statistice din fiecare. Poligonul frecvențelor prezintă și el un avantaj în fața histogramei, dându-ne posibilitate să comparăm direct două sau mai multe distribuții (figura 2.3).

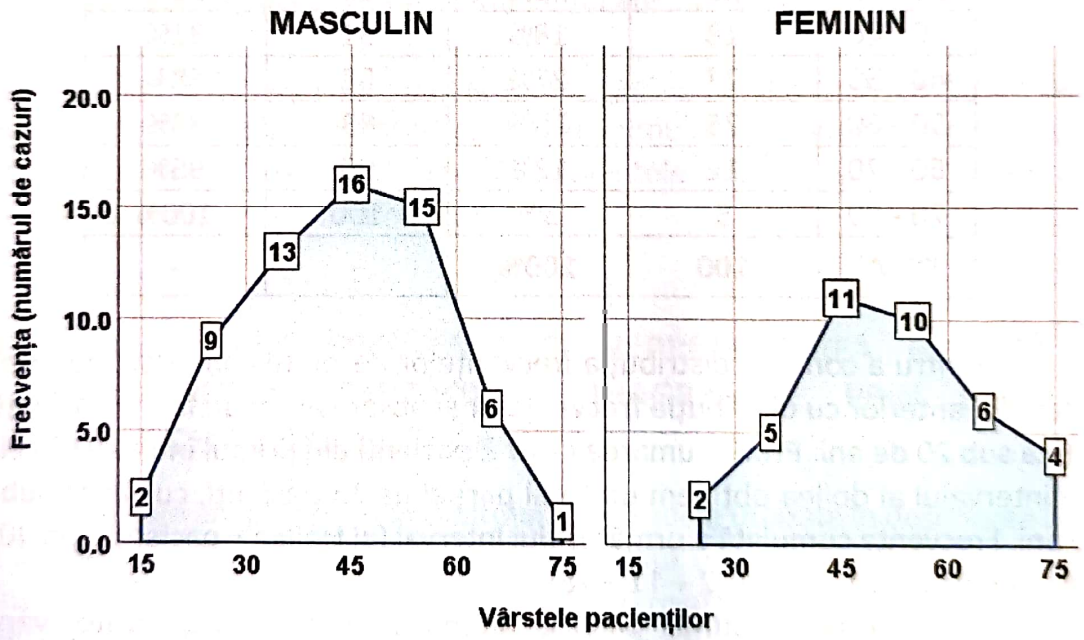


Figura 2.3. Poligoanele frecvențelor vârstelor, separat pentru cele două genuri: masculin și feminin

De exemplu, dacă se dorește realizarea unei comparații între distribuțiile vârstelor bărbaților și femeilor din eșantionul de 100 de pacienți (figura 2.3), observăm faptul că vârsta tipică (întâlnită cel mai des) este relativ similară între cele două grupuri, situându-se în jurul valorii de 45 de ani. De asemenea, nivelul variabilității (împrăștierii) este puțin mai ridicat în situația genului masculin (15 –

Scanned with OKEN Scanner

75) ani, comparativ cu genul feminin (25 – 75) ani. Tot din analiza graficului prezentat în figura 2.3, se observă o ușoară predominanță a persoanelor cu vârsta sub 45 de ani în cazul genului masculin și a celor peste acest nivel de vârstă în cazul genului feminin.

Continuând pe linia exemplului anterior, privind distribuția pacienților în funcție de variabila vârstă, presupunem faptul că am fi interesați în cunoașterea numărului de pacienți cu vârsta mai mică de 30 de ani. Aceste valori pot fi aproximate prin construirea unui distribuții a frecvențelor cumulate (tabelul 2.5) și, ulterior, prin reprezentarea grafică a acesteia în forma poligonului frecvențelor absolute cumulate.

Tabelul 2.5. Distribuția frecvențelor cumulate
în cazul variabilei „Vârstă"

Vârsta	Frecvența		Frecvența cumulată	
	absolută	relativă	absolută	relativă
10 - 20	2	2%	2	2%
20 - 30	11	11%	13	13%
30 - 40	18	18%	31	31%
40 - 50	27	27%	58	58%
50 - 60	25	25%	83	83%
60 - 70	12	12%	95	95%
70 - 80	5	5%	100	100%
TOTAL	**100**	**100%**	-	-

Pentru a construi distribuția frecvențelor absolute cumulate, pornim de la tabelul anterior cu distribuția frecvențelor și observăm că au fost 2 pacienți cu vârsta sub 20 de ani. Prin însumarea celor 2 pacienți din primul interval cu cei 11 din intervalul al doilea obținem un total parțial de 13 pacienți, cu vârsta sub 30 de ani. Frecvența cumulată a următorului interval (al treilea – pacienții sub 40 de ani) care urmează este 31 (2 + 11 + 18).

Acest proces continuă până la ultimul interval, corespunzător vârstei maxime. Distribuția frecvențelor relative cumulate se realizează similar, cumulând din aproape în aproape frecvențele relative ale intervalelor de variație, în ordinea crescătoare a valorii intervalelor. De exemplu, frecvența relativă cumulată de 95% pentru penultimul interval ne indică faptul că 95% dintre pacienți au vârsta sub 70 de ani.

În vederea reprezentării grafice, se vor uni punctele având abscisa limita superioară a fiecărui interval și ordonata frecvențele cumulate corespunzătoare din tabel (figura 2.4). Limitei inferioare a primului interval i se atribuie

Scanned with OKEN Scanner

convențional valoarea zero (în exemplul nostru, niciun pacient nu are vârsta sub 10 ani).

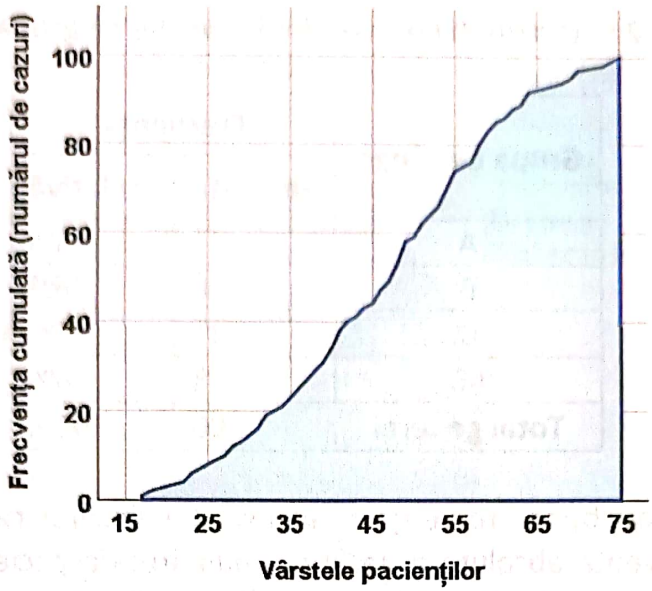


Figura 2.4. Poligonul frecvențelor cumulate ale vârstelor pacienților

Poligonul frecvențelor cumulate are exact aceeași formă, indiferent dacă utilizăm în reprezentarea grafică frecvențele cumulate absolute sau relative, cu precizarea că în situația când folosim frecvențele relative, valorile de pe axa verticală vor fi exprimate în procente.

2.3. TABELELE DE FRECVENȚE ȘI REPREZENTAREA GRAFICĂ A VARIABILELOR CALITATIVE - DIAGRAMELE PRIN BARE ȘI GRAFICELE DE STRUCTURĂ

Așa cum am prezentat în capitolul 1, metodele utilizate în descrierea unui set de date aparțin statisticii descriptive. Aceasta organizează datele statistice pentru a surprinde trăsăturile esențiale ale lor, pentru a descoperi tendința de concentrare a datelor și a identifica valorile extreme, neobișnuite ale variantelor pe care le poate lua o variabilă statistică.

Tabelul de frecvențe se realizează prin gruparea tuturor variantelor unei variabile calitative în mai multe categorii (clase) distincte, evidențiind numărul de observații aferent fiecărei categorii în parte.

În setul de date prezentat anterior în tabelul 2.1, sunt evidențiate șapte variabile statistice (vârsta pacienților, genul, grupa de sânge, glicemia, colesterolul, HTA și diabet). Variabilele statistice 2, 4 și 5 sunt cantitative (numerice), iar celelalte patru sunt calitative (categoriale).

Scanned with OKEN Scanner

Vom realiza tabelul frecvențelor absolute pentru cei 100 de pacienți, pentru variabila statistică "grupa de sânge"(tabelul 2.6).

Tabelul 2.6. Distribuția pacienților în funcție de grupa de sânge

Grupa de sânge	Frecvența	
	absolută	relativă
A	25	25%
B	36	36%
0	31	31%
AB	8	8%
Total general	**100**	**100%**

Pentru a obține frecvențele relative ale fiecărei categorii (clase), se raportează frecvența absolută a fiecăreia (numărul de pacienți pentru fiecare grupă de sânge) la numărul total de observații (numărul tuturor al pacienților din eșantion). Frecvența relativă conturează legătura fiecărei clase de frecvență cu numărul total de observații (ponderea sau procentul deținut de acestea din total).

În practică, pentru reprezentarea grafică a datelor de tip calitativ (categoriale) se utilizează cel mai frecvent diagramele prin bare și diagramele de structură, de obicei sub formă circulară.

În cazul *diagramelor prin bare*, tipurile calitative (categoriile) sunt prezentate pe axa orizontală, iar clasele de frecvențe pe axa verticală, fiind proporționale cu înălțimea fiecărei bare (figura 2.5).

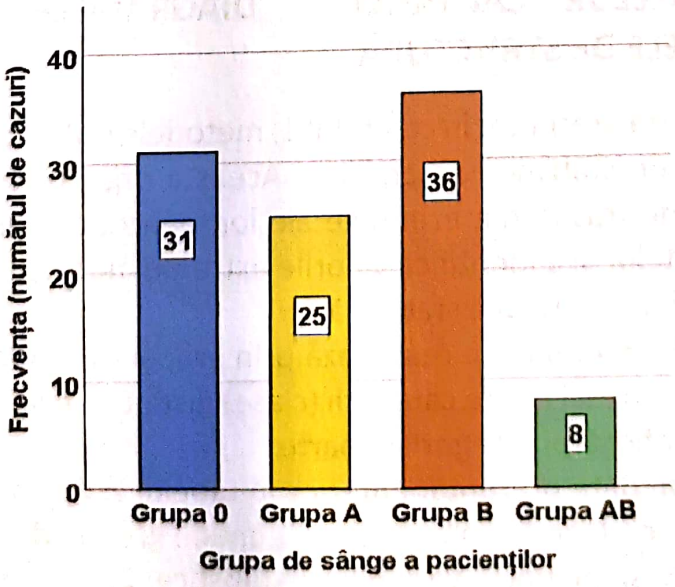


Figura 2.5. Diagrama prin bare a grupelor de sânge ale pacienților

Scanned with OKEN Scanner

Variabila „grupa de sânge" este măsurată pe o scară nominală, deci ordinea celor 4 grupe de sânge pe axa orizontală nu este importantă. Înălțimea barelor corespunde numărului de pacienți care au o anumită grupă de sânge.

Diagramele de structură evidențiază proporția (procentul) pe care fiecare categorie îl deține în volumul total al colectivității studiate (figura 2.6).

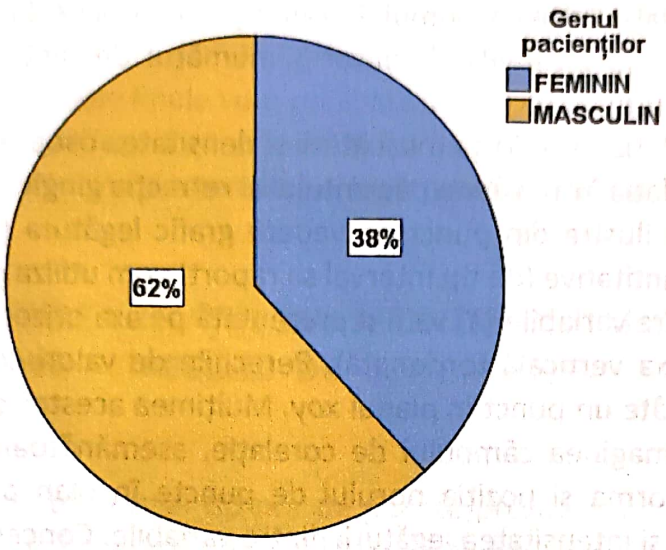


Figura 2.6. Diagrama de structură pentru genul pacienților

Deoarece, sectoarele de cerc reprezintă frecvența relativă a fiecărui gen (Masculin / Feminin), ca procent din numărul total al pacienților, se poate realiza cu ușurință o comparație între ponderile deținute de cele două sexe. Din analiza diagramei se remarcă imediat faptul că eșantionul supus analizei este format în principal din persoane de sex masculin (în proporție de 62%).

2.4. PREZENTAREA RELAȚIEI SAU ASOCIERII DINTRE DOUĂ VARIABILE CANTITATIVE - DIAGRAMA TIP „NOR DE PUNCTE"

Anterior am prezentat metodele grafice de centralizare a distribuțiilor statistice pentru o singură variabilă sau distribuțiilor univariate. Există însă situații în care dorim să studiem și să realizăm o reprezentare grafică a relației (legăturii) dintre două variabile statistice, situație în care analizăm distribuții statistice bivariate.

În acest sens, ne putem gândi la următoarele exemple de posibile legături sau corelații între două variabile cantitative (numerice):

- un antropolog dorește să studieze corelația sau legătura potențială dintre înălțimea taților și cea a copiilor acestora; cu alte cuvinte, tații mai înalți tind să aibă de asemenea și copii mai înalți?

Scanned with OKEN Scanner

- un medic dorește să investigheze posibila corelație dintre indicele masei corporale a pacienților și valorile glicemiei à jeun ale acestora;
- legătura dintre vârsta pacienților și tensiunea arterială sistolică a lor;
- corelația între volumul expirator forțat într-o secundă (FEV1) și numărul de ani în care pacienții au fumat;
- corelația între nivelul colesterolului LDL și grosimea plăcii de aterom;
- corelația între consumul de zahăr și numărul de carii dentare;
- corelația între pH-ul salivar și numărul de dinți cariați, lipsă sau obturați;
- corelația între forța mușcăturii și densitatea osoasă mandibulară;
- corelația între vârsta pacientului și retracția gingivală, etc.

Pentru a ilustra din punct de vedere grafic legătura (asocierea) dintre două variabile cantitative (de tip interval sa raport) vom utiliza *graficul tip nor de puncte*. Una dintre variabile (*X*) va fi reprezentată pe axa orizontală (abscisă), iar a doua (*Y*) pe axa verticală (ordonată). Perechile de valori corelate (x_i, y_i) se reprezintă prin câte un punct în planul *xoy*. Mulțimea acestor puncte dispersate în grafic oferă imaginea câmpului de corelație, asemănătoare cu un „nor de puncte". După forma și poziția norului de puncte în plan poate fi apreciată existența, forma și intensitatea legăturii dintre variabile. Concentrarea punctelor în jurul unei anumite drepte sau curbe ne va sugera faptul că între variabile există o legătură. Inexistența unei legături între cele 2 variabile poate fi apreciată în situația împrăștierii uniforme a tuturor punctelor în spațiul *xoy* sau în situația când punctele se concentrează în jurul unei drepte care este paralelă cu una dintre cele două axe.

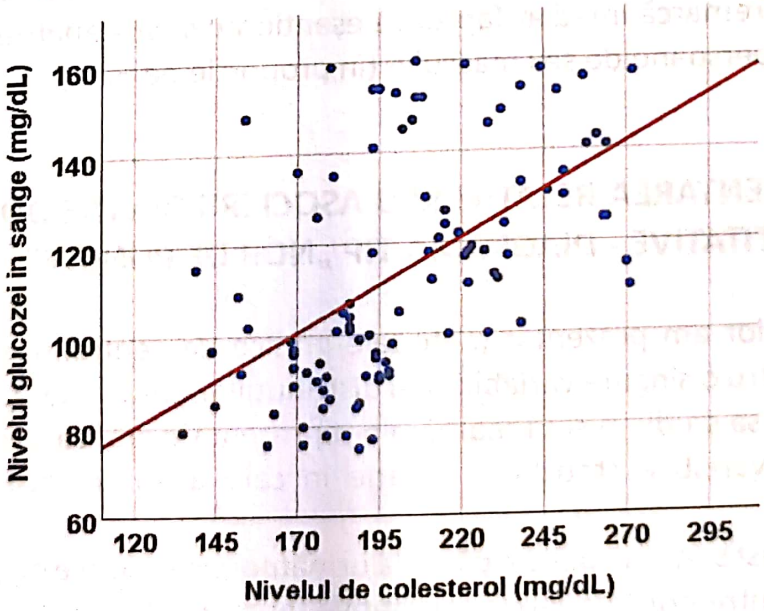


Figura 2.7. Corelația dintre valorile colesterolului și ale glicemiei

Scanned with OKEN Scanner

Revenind la exemplul prezentat inițial, colesterolul pacienților reprezintă una dintre variabile, iar glicemia cealaltă variabilă analizată (tabelul 2.1).

Există vreo legătură între nivelul colesterolului și nivelul glicemiei pentru pacienții din eșantion? Poate fi rezonabil să concluzionăm că un colesterol mai mare este asociat cu o valoare mai ridicată a glicemiei?

Diagrama construită conturează o legătură de intensitate medie între aceste două variabile statistice, care pare să evidențieze o relație sau o asociere între nivelul colesterolului și cel al glicemiei pacienților din eșantionul studiat (figura 2.7). În capitolele finale vom prezenta aceste corelații nu doar prin prisma reprezentării lor grafice ci și prin calculul unor indicatori statistici care să contureze mai clar forma și intensitatea legăturii dintre variabile.

2.5. PREZENTAREA RELAȚIEI SAU ASOCIERII DINTRE DOUĂ VARIABILE CALITATIVE - TABELELE DE CONTINGENȚĂ ȘI GRAFICUL PRIN COLOANE

Realizarea diagramei tip nor de puncte presupune că ambele variabile analizate trebuie să fie numerice, cel puțin de tip interval. În situația în care dorim să studiem relația dintre două variabile statistice calitative (de tip nominal sau ordinal) vom centraliza datele sub forma unui *tabel de contingență*, iar reprezentarea grafică se va realiza sub forma unui model de tip stivă de bare, în care vor fi evidențiate în vederea comparării procentele deținute de fiecare categorie (numărul de observații care posedă simultan un anumit nivel al fiecărei variabile).

Pentru a exemplifica, vom utiliza aceleași date, încercând să realizăm o comparație pentru a identifica o posibilă asociere între variabilele nominale, dihotomice (cu doar 2 categorii) "HTA" și "Diabet". Așa cum am precizat anterior, realizarea unui tabel de contingență presupune doar utilizarea unor variabile calitative, de tip nominal sau ordinal.

Deoarece una dintre variabilele analizate este de tip cantitativ (Glicemia), vom realiza o conversie într-o variabila calitativă împărțind valorile numerice în două mari categorii, mai mari sau mai mici decât nivelul prag de 110 mg / dL.

În consecință, variabila continua "Glicemie" a fost transformată într-o nouă variabilă "Diabet", având doar două variante posibile DA sau 1 pentru persoanele cu nivelul glicemiei peste 110 si NU sau 0 pentru pacienții având glicemia sub acest nivel prag. Variabila HTA a fost prezentată sub forma unei variabile calitative având două variante posibile DA sau 1 pentru pacienții cu hipertensiune și NU sau 0 pentru cei fără. Pe coloanele tabelului de contingență am așezat cele 2 categorii ale variabilei "Diabet", iar pe linii cele 2 categorii ale variabilei "HTA" (tabelul 2.7). Această variantă de prezentare a informațiilor ne

Scanned with OKEN Scanner

oferă posibilitatea realizării unei comparații a pacienților cu / fără diabet pentru ambele variante ale variabilei HTA (cu / fără hipertensiune).

Tabelul 2.7. Tabelul de contingență pentru verificarea asocierii între variabilele HTA și Diabet

HTA \ DIABET	DA	NU	Total
NU	24	34	**58**
DA	27	15	**42**
Total	**51**	**49**	**100**

Pentru a vizualiza mai clar potențiala asociere între aceste două variabile categoriale vom utiliza *graficul prin coloane* (figura 2.8)

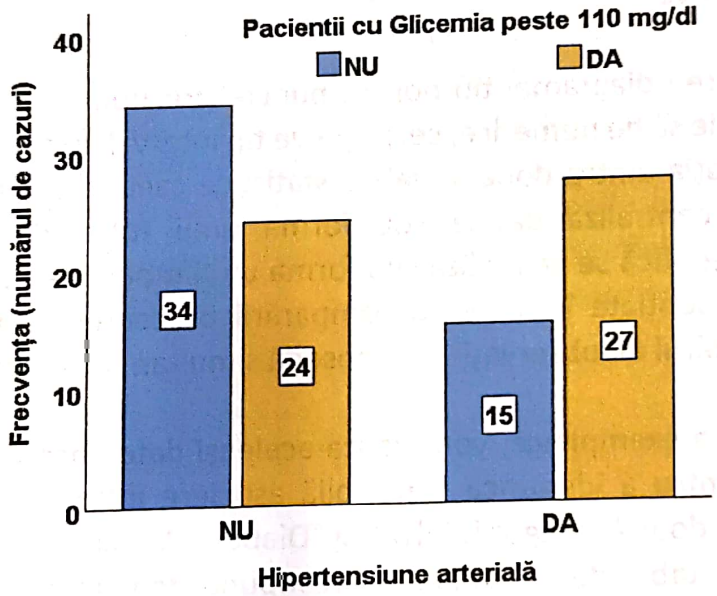


Figura 2.8. Verificarea grafică a asocierii dintre 2 variabile categoriale (HTA și glicemia)

Remarcăm faptul că dintre pacienții care au hipertensiune arterială 27 (64.3%) au și diabet zaharat, iar dintre pacienții fără hipertensiune arterială 34 (58.6%) nu au nici diabet. Astfel, cei mai mulți pacienți cu hipertensiune arterială au și diabet și, la polul opus, dintre pacienții fără hipertensiune arterială, majoritatea nu au nici diabet.

Realizând această analiză descriptivă preliminară, atât prin centralizarea datelor sub forma unui tabel de contingență (cu dublă intrare) cât și prin

Scanned with OKEN Scanner

reprezentarea grafică aferentă, se conturează o posibilă asociere între cele două variabile statistice calitative studiate "HTA" și "Diabet".

2.6. PREZENTAREA RELAȚIEI SAU ASOCIERII DINTRE O VARIABILĂ CANTITATIVĂ ȘI UNA CALITATIVĂ - PIRAMIDA VÂRSTELOR

Reprezentarea grafică a structurii pe vârste și sexe a unei populații se poate face cu ajutorul unui grafic special numit piramida vârstelor, care este o combinare a două histograme pe vârste, câte una pentru fiecare sex, având scările inversate. Cele două abscise se înscriu pe verticală, iar ordonatele pe orizontală: în stânga pentru sexul masculin, iar în dreapta sexul feminin.

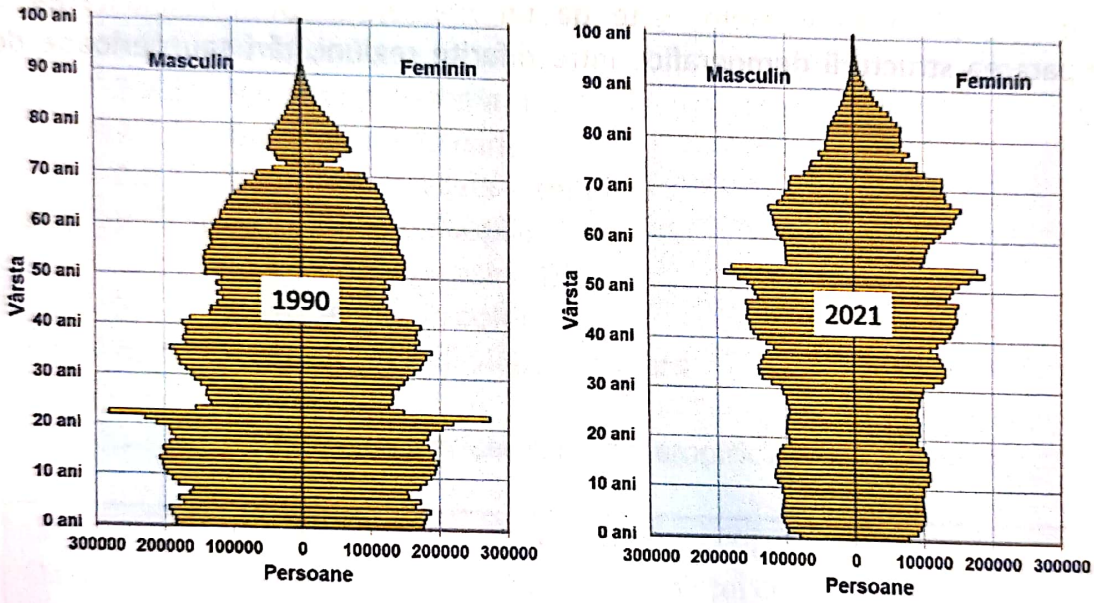


Figura 2.9. Piramidele vârstelor populației României pentru anii 1990 și 2021
Sursa: Institutul Național de Statistică http://statistici.insse.ro:8077/tempo-online/#/pages/tables/insse-table

Efectivele diferitelor generații ale celor două sexe se reprezintă prin benzi orizontale de lungime proporțională, obținând-se în final un grafic în formă apropiată de un triunghi isoscel, de unde derivă și denumirea de piramidă a vârstelor.

Piramida vârstelor permite evidențierea dublului efect al vârstei și al generației asupra evoluției în timp a populațiilor: odată cu înaintarea în vârstă, efectivul generațiilor se micșorează în mod progresiv sub efectul vârstei, fapt care explică forma triunghiulară a piramidei; în același timp, fiecare generație are propria sa istorie, adică efectul generației (diferențele care există între efective la naștere, între nivelul mortalității diferențiat pe sexe de-a lungul întregii vieți, de

Scanned with OKEN Scanner

migrații diferențiate etc.) explică intrândurile și proeminențele ce apar pe laturile piramidei

Piramida vârstelor nu este perfect simetrică pentru cele două sexe. În prima jumătate a vieții efectivele generațiilor masculine sunt mai mari decât cele feminine, ca urmare a supra-natalității și a supra-mortalității masculine. În anii de mijloc ai vieții efectivele generațiilor sunt sensibil egale, iar apoi se înregistrează, în partea a doua a vieții, o preponderență a efectivelor feminine, tot mai evidentă pe măsura înaintării în vârstă, ca urmare a supramortalității masculine la toate vârstele și cu deosebire la vârstele înaintate.

Această metodă grafică ajută la identificarea rapidă a tendințelor fenomenelor sociale, economice și politice cum ar fi îmbătrânirea demografică a populației, scăderea natalității, supra-mortalitatea masculină, impactul războaielor, migrațiile importante, crizele economice sau epidemiile. De asemenea, piramida vârstelor este de un real ajutor pentru analizarea și compararea structurii demografice între diferite regiuni, țări sau perioade de timp.

Scanned with OKEN Scanner

Capitolul. 3

INDICATORII TENDINȚEI CENTRALE, VARIAȚIEI ȘI FORMEI DISTRIBUȚIILOR

SUBIECTE:

3.1. Indicatorii tendinței centrale
- 3.1.1. Media aritmetică
- 3.1.2. Medii cu aplicații speciale
- 3.1.3. Mediana
- 3.1.4. Modul

3.2. Indicatorii variației sau dispersiei
- 3.2.1. Amplitudinea variației
- 3.2.2. Varianța și deviația standard
- 3.2.3. Coeficientul de variație
- 3.2.4. Cuartilele, decilele și centilele

3.3. Indicatorii formei distribuțiilor
- 3.3.1. Coeficientul de asimetrie
- 3.3.2. Coeficientul de boltire

3.4. Importanța practică a deviației standard
- 3.4.1. Scorurile z
- 3.4.2. Teorema lui Cebîșev și regula empirică

La finalizarea acestui capitol, studentul va fi capabil să:
O3-1: calculeze și interpreteze indicatorii tendinței centrale (media aritmetică, media armonică, media geometrică, media pătratică, mediana și modul)
O3-2: calculeze și interpreteze indicatorii variației (amplitudinea variației, varianța și deviația standard)
O3-3: să explice și să aplice corect teorema lui Cebîșev și regula empirică
O3-4: să determine și să interpreteze cuartilele, decilele și centilele
O3-5: să calculeze și interpreteze coeficienții de asimetrie și de boltire
O3-6: să determine și să ofere o interpretare practică a scorurilor z

3.1. INDICATORII TENDINȚEI CENTRALE

În capitolul anterior am văzut modalitățile de centralizare a datelor statistice într-o formă clară, sugestivă și ușor de interpretat și ulterior de

Scanned with OKEN Scanner

prezentare a lor sub formă grafică, indiferent dacă variabilele statistice studiate îmbracă o formă calitativă sau cantitativă.

În continuare, vom prezenta trei modalități numerice (indicatori statistici) pentru a descrie variabilele cantitative și anume indicatorii tendinței centrale sau mărimile medii, indicatorii dispersiei sau împrăștierii și indicatorii formei distribuțiilor statistice.

O mărime medie, în general, reprezintă un indicator al localizării care evidențiază valoarea centrală a unei variabile statistice sau mijlocul unei distribuții.

Exemple de mărimi medii întâlnite în practică:

- *vârsta medie a populației rezidente* a României s-a majorat de la 40,8 ani în 2012 până la 42,3 ani în 2021;
- *vârsta medie a mamei la prima naștere* în România a crescut de la 22,3 ani în 1990 până la 27,7 ani în 2021;
- *greutatea medie la naștere* a unui bebeluș născut la termen (săptămânile 37-42) este de aproximativ 3,4 kg;
- *înălțimea medie la naștere* a unui bebeluș născut la termen (săptămânile 37-42) este de aproximativ 49 cm;
- *numărul mediu de pensionari* a crescut de la 3.679.000 persoane în anul 1990 la 5.079.000 persoane în anul 2021 (cu aproximativ 38%).

Într-o cercetare statistică completă imaginea generală conturată prin reprezentarea grafică a datelor este insuficientă, fiind preferată determinarea unui indicator de sinteză, care să reprezinte nivelul general al fenomenului studiat, modelul lui central de dezvoltare – respectiv *o valoare reprezentativă*, în jurul căreia are loc o tendință de concentrare a nivelurilor individuale ale variabilei studiate.

Fenomenele colective supuse analizei statistice sunt purtătoare a mai multor caracteristici, fiecare cu un număr mare de variante ca urmare a îmbinării influențelor diferiților factori sub acțiunea cărora iau naștere și se dezvoltă aceste fenomene. Variația unei caracteristici în cadrul unei colectivități este determinată de acțiunea factorilor esențiali asociați cu factorii neesențiali sau întâmplători.

Factorii esențiali acționează asupra tuturor unităților dintr-o colectivitate și determină dezvoltarea lor normală sub aspectul unei anumite caracteristici sau tendința normală, centrală de dezvoltare a unei colectivități. Asupra fenomenelor acționează și factori întâmplători (neesențiali), care determină abateri în ambele sensuri ale nivelurilor individuale ale unei variabile de la tendința normală de dezvoltare.

Mărimile medii evidențiază deci *influența factorilor esențiali* care au acționat asupra unităților statistice dintr-o colectivitate sub aspectul unei caracteristici care prezintă interes în cercetare. Influenta factorilor neesențiali sau întâmplători, care determină împrăștierea nivelurilor individuale ale unei

Scanned with OKEN Scanner

variabile în jurul nivelului ei mediu, este măsurată de indicatorii variației (dispersiei).

Esența mărimilor medii constă în faptul că ele caracterizează în mod sintetic ceea ce este esențial, normal, tipic și reprezentativ în cadrul unei colectivități statistice sub aspectul unei anumite variabile de interes (valori tip). Cu ajutorul mediilor se pot compara colectivitățile de fenomene și se pot desprinde anumite trăsături generale asupra nivelului lor de dezvoltare. Pentru a se asigura mediei un conținut real și reprezentativ, colectivitățile studiate trebuie să fie *omogene*.

Media reprezintă *o valoare teoretică a variabilei* (spre care tind toate nivelurile ei individuale) și se exprimă în aceleași unități de măsură ca și aceasta.

Proprietățile mărimilor medii:

- să fie definite precis și obiectiv;
- să posede o semnificație concretă;
- să poată fi calculate cu rapiditate;
- trebuie să se preteze cu ușurință la calcule algebrice ulterioare (mai ales în studiul comparativ al mai multor serii statistice).

Principalele mărimi medii utilizate în statistică sunt: media aritmetică, media armonică, media geometrică, media pătratică, mediana și modul.

3.1.1. MEDIA ARITMETICĂ

Media aritmetică a unui șir de valori individuale ale unei variabile statistice este egală cu raportul dintre suma și numărul lor (aceasta este influențată mai puternic de valorile mari ale variabilei analizate). Media poate fi calculată utilizând toate variantele variabilei studiate de la nivelul populației generale sau doar pentru informațiile de la nivelul unui eșantion reprezentativ extras din aceasta.

Media aritmetică a unei variabile X cu variantele ei $x_1, x_2, \dots, x_i, \dots, x_n$, în cazul unui eșantion de volum n extras dintr-o populație generală de volum N, reprezintă o *statistică*:

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_i + \cdots + x_n)}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{3.1}$$

Media aritmetică a unei variabile X cu variantele ei $x_1, x_2, \dots, x_i, \dots, x_N$, în cazul unei populații de volum N reprezintă un *parametru*:

$$\mu = \frac{(x_1 + x_2 + \cdots + x_i + \cdots + x_N)}{N} = \frac{\sum_{i=1}^{N} x_i}{N} \tag{3.2}$$

Scanned with OKEN Scanner

În anumite situații particulare, o variabilă poate înregistra una sau mai multe valori neobișnuit de mari sau de mici, care să influențeze într-un mod nedorit valoarea mediei, astfel încât aceasta să nu mai poată servi drept indicator corespunzător al tendinței centrale. Pentru a elimina efectul negativ al acestor valori extreme putem înlătura un procent redus (de exemplu 5%) al acestora din cadrul setului de date. Media valorilor rămase ale variabilei studiate poartă denumirea de *medie aritmetică ajustată*.

Proprietățile mediei aritmetice:

- variabilele utilizate în calculul mediei trebuie să fie *numerice*, de tip interval sau raport (indicatorul nu are sens în cazul variabilelor calitative);
- în calculul mediei aritmetice sunt incluse *toate variantele variabilei*;
- media aritmetică este *unică* pentru fiecare variabilă statistică în parte, fiind cuprinsă între valorile minimă și maximă ale acesteia;
- media aritmetică a unui șir de valori constante este egală cu valoarea lor comună;
- media aritmetică este *asociativă* (dacă se înlocuiesc mai multe variante ale variabilei cu media lor, atunci media seriei modificate este egală cu media seriei inițiale);
- media este *translativă* (dacă se mărește sau se micșorează fiecare nivel individual al variabilei cu o cantitate constantă, atunci media se mărește sau se micșorează cu acea constantă);
- *suma tuturor abaterilor fiecărei valori individuale a unei variabile de la nivelul mediu al acesteia este zero:* $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

Tabelul 3.1. Ilustrarea proprietăților mediei aritmetice

Persoana	Înălțimea x_i	Media $\bar{x}$	Diferențele $x_i - \bar{x}$	Înălțimea – 10 cm $x_i - 10$
1	170	176	-6	160
2	177	176	1	167
3	175	176	-1	165
4	185	176	9	175
5	182	176	6	172
6	168	176	-8	158
7	172	176	-4	162
8	178	176	2	168
9	166	176	-10	156
10	187	176	11	177
Total	**1760**	**1760**	**0**	**1660**
Media	**176**	**176**	**0**	**166**

Scanned with OKEN Scanner

În continuare am evidențiat aplicarea proprietăților mediei aritmetice în cadrul unui set de date fictive privind variabila înălțime (X), pentru un eșantion format din 10 persoane (tabelul 3.1). Din analiza tabelului 3.1 observăm faptul că, media înălțimii celor 10 persoane este 176 cm, fiind cuprinsă între nivelul minim (166 cm) și cel maxim (187 cm). Presupunând că toate cele 10 persoane ar avea aceeași înălțime (de exemplu 176 cm), media seriei modificate ar fi tot 176 cm (egală cu valoarea comună). De asemenea, dacă considerăm ipotetic faptul că toate persoanele ar avea înălțimea mai mică cu 10 cm decât în seria inițială, noua medie calculată pentru aceste valori diminuate ar fi și ea cu 10 cm mai redusă (166 cm). În final, dacă din fiecare variantă a variabilei înălțime scădem valoarea medie (coloana 4 a tabelului 3.1), remarcăm faptul că, abaterile cu semnul minus se compensează cu abaterile pozitive, suma tuturor acestor diferențe fiind zero.

În situația în care variabila statistică X, care intră în calculul mediei, prezintă mai multe observații x_i care au aceeași valoare n_i (ponderi sau frecvențe de apariție), vom calcula *media aritmetică ponderată*:

- pentru un eșantion de volum n:

$$\bar{x} = \frac{x_1 \cdot n_1 + x_2 \cdot n_2 + \cdots + x_n \cdot n_n}{n_1 + n_2 + \cdots + n_n} = \frac{\sum_{i=1}^{n} x_i \cdot n_i}{\sum_{i=1}^{n} n_i} \tag{3.3}$$

- pentru o populație de volum N:

$$\mu = \frac{\sum_{i=1}^{N} x_i \cdot n_i}{\sum_{i=1}^{N} n_i} \tag{3.4}$$

Exemplu: o farmacie a vândut 10 medicamente, 3 din sortimentul A cu prețul unitar de 125 lei, 5 din sortimentul B cu prețul unitar de 95 lei și 2 din sortimentul C cu prețul unitar de 145 lei. Care este prețul mediu de vânzare ale celor 10 medicamente vândute:

$$\bar{x} = \frac{125 + 125 + 125 + 95 + 95 + 95 + 95 + 95 + 145 + 145}{10} =$$

$$= \frac{3 \cdot 125 + 5 \cdot 95 + 2 \cdot 145}{10} = \frac{1140}{10} = 114\ lei$$

În cazul unei serii cu frecvență având variabila distribuită pe intervale este necesar să se determine mai întâi câte o valoare reprezentativă a fiecărui interval. Admițându-se convențional că, în cadrul fiecărui interval, variabila studiată se distribuie în mod liniar uniform, mijlocul intervalului respectiv x'_i este valoarea reprezentativă a fiecărui interval. În această situație vom aplica formulele mediei

Scanned with OKEN Scanner

ponderate prezentate anterior, utilizând mijloacele intervalelor de variație ale variabilei statistice.

În cadrul formulei de calcul a mediei aritmetice ponderate se pot utiliza atât frecvențele absolute n_i sau cele relative f_i, rezultatul final fiind identic.

3.1.2. MEDII CU APLICAȚII SPECIALE

a) *Media armonică* este o formă transformată a mediei aritmetice și se determină ca o medie aritmetică inversă, calculată din mărimile inverse ale variantelor variabilei (se utilizează în calculul valorii medii în cazul mărimilor relative, fiind mai puternic influențată de valorile mici ale variabilei). Formula se poate aplica numai atunci când toate variantele variabilei studiate sunt valori pozitive ($x_i \neq 0$).

- pentru un eșantion de volum n:

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \tag{3.5}$$

- pentru o populație de volum N:

$$\bar{\mu}_h = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}} \tag{3.6}$$

Exemplul 1: Presupunem că un anumit medicament își modifică prețul în timpul unui trimestru de 3 ori. Astfel, în prima lună cu 500 de lei se puteau cumpăra 26 cutii, în a două lună cu aceeași sumă 22 cutii, iar în cea de a treia doar 18 cutii. Se cere să se determine cantitatea medie vândută și prețul mediu al acestui medicament pentru întreg trimestrul considerat.

Dacă am aplica media aritmetică: $\bar{x} = (26 + 22 + 18)/3 = 22\ cutii$, iar prețul mediu ar fi: $500/22 = 22{,}7\ lei/cutie$. Acest preț mediu nu ține însă cont de cantitățile reale ale medicamentului care au fost vândute în fiecare lună a trimestrului. Dacă am cunoaște aceste cantități, atunci cantitatea medie și prețul mediu ar putea fi determinate prin aplicarea mediei aritmetice ponderate.

Determinarea corectă a numărului mediu de cutii vândute se poate realiza cu ajutorul mediei armonice:

$$\overline{x_h} = \frac{3}{\sum_{i=1}^{3} \frac{1}{x_i}} = \frac{3}{\frac{1}{26} + \frac{1}{22} + \frac{1}{18}} = 21{,}51\ cutii$$

Scanned with OKEN Scanner

Deci, numărul mediu de cutii care se puteau cumpăra în timpul trimestrului cu 500 de lei a fost de doar 21,51 cutii (cu jumătate de cutie mai mic decât am obținut prin aplicarea mediei aritmetice inițial), iar prețul mediu pentru o cutie de medicamente a fost de aproximativ $500/21{,}51 = 23{,}25\ lei/cutie$.
Verificare: prețul mediu de vânzare în prima lună a fost de $500/26 = 19{,}23\ lei/cutie$, în a doua lună $500/22 = 22{,}72\ lei/cutie$ și în a treia lună $500/18 = 27{,}78\ lei/cutie$. Calculând media aritmetică a acestor 3 prețuri medii lunare obținem prețul mediu de vânzare pentru întreg trimestrul de 23,25 lei / cutie.

În situația în care variabila statistică X, care intră în calculul mediei, prezintă mai multe observații x_i care au aceeași valoare n_i (ponderi sau frecvențe de apariție), vom calcula *media armonică ponderată*:

- pentru un eșantion de volum n:

$$\bar{x}_h = \frac{\sum_{i=1}^{n} n_i}{\sum_{i=1}^{n} \frac{n_i}{x_i}} \tag{3.7}$$

- pentru o populație de volum N:

$$\bar{\mu}_h = \frac{\sum_{i=1}^{N} n_i}{\sum_{i=1}^{N} \frac{n_i}{x_i}} \tag{3.8}$$

Exemplul 2: Presupunem că ratele de mortalitate (numărul de decese la 1000 de locuitori) în patru localități sunt 8‰, 9‰, 11‰ și 12 ‰. Populațiile celor 4 localități sunt 15000 locuitori, 18000 locuitori, 20000 locuitori, respectiv 30000 locuitori. Vom determina rata medie a mortalității pentru regiunea formată din cele 4 localități:

$$\bar{x}_h = \frac{\sum_{i=1}^{4} n_i}{\sum_{i=1}^{4} \frac{n_i}{x_i}} = \frac{15000 + 18000 + 20000 + 30000}{\frac{1}{8} \cdot 15000 + \frac{1}{9} \cdot 18000 + \frac{1}{11} \cdot 20000 + \frac{1}{12} \cdot 30000}$$

Efectuând calculele obținem rata medie a mortalității: $\bar{x}_h = 10{,}13\ ‰$

b) În cazul fenomenelor care se dezvoltă în progresie geometrică, pentru determinarea nivelului lor mediu, se utilizează *media geometrică*. Această mărime medie este folosită de exemplu în calcularea ritmului mediu de creștere anuală sau în dinamica sporului natural al populației (este influențată mai puternic de valorile mici ale șirului).

Scanned with OKEN Scanner

- pentru un eșantion de volum n:

$$\bar{x}_g = \sqrt[n]{\prod_{i=1}^{n} x_i} \tag{3.9}$$

- pentru o populație de volum N:

$$\bar{\mu}_g = \sqrt[N]{\prod_{i=1}^{N} x_i} \tag{3.10}$$

Exemplu: dacă o tulpină de bacterii își mărește populația cu 20% în prima oră, 30% în următoarea oră și 50% în a treia oră, putem calcula o valoare medie orară estimată a creșterii procentuale a populației acestora.
Vom utiliza în acest caz media geometrică în locul mediei aritmetice.

Să presupunem că avem la început o populație de 100 de bacterii, care se va majora pe parcursul primei ore la 120 de bacterii ($= 100 + 100 * 20\% = 100 * 120\%$). După cea de a doua oră, numărul lor ajunge la 156 de bacterii ($= 120 + 120 * 30\% = 120 * 130\%$), iar la sfârșitul celei de a treia ore populația de bacterii va fi de 234 unități ($= 156 + 156 * 50\% = 156 * 150\%$).

Tabelul 3.2. Creșterea populației de bacterii în progresie geometrică

Timp	Creștere orară		Nr. bacterii (inițial 100)	Creștere medie orară ($\bar{x}_g$)		Nr. bacterii (inițial 100)
	%	coef.		%	coef.	
1 oră	120%	1,20	120	132,76%	1,33	133
2 ore	130%	1,30	156	132,76%	1,33	176
3 ore	150%	1,50	234	132,76%	1,33	234

Calculăm media geometrică a acestor creșteri:

$$\bar{x}_g = \sqrt[3]{1{,}2 \cdot 1{,}3 \cdot 1{,}5} = 1{,}3276 \text{ sau } 132{,}76\%$$

Astfel, rata medie orară de creștere a populației de bacterii pe o perioadă de 3 ore este de 32,76% (dacă ar fi crescut uniform pe parcursul celor 3 ore, atunci pornind de la 100 de bacterii am ajunge la 234). Calculele detaliate sunt prezentate în tabelul 3.2.

Scanned with OKEN Scanner

În concluzie, de fiecare dată când ne confruntăm cu o creștere procentuală a unui fenomen într-o perioadă de timp, rata medie de variație se va determina prin aplicarea formulei mediei geometrice.

c) *Media pătratică* se folosește în situația în care fenomenele înregistrează creșteri, aproximativ, după o lege exponențială (creșterea este mai lentă la începutul seriei de date statistice și din ce în ce mai pronunțată spre sfârșitul acesteia), fiind utilizată în analiza tendințelor de evoluție neliniare, de tip exponențial (este influențată mai puternic de valorile mari ale șirului). Media pătratică este folosită și ca model matematic in calculul indicatorilor sintetici ai variației, de exemplu în determinarea deviației standard.

În cazul unei variabile statistice X, având variantele $x_1, x_2, x_3, \ldots, x_i, \ldots, x_n$ media pătratică se determină după formula:

- pentru un eșantion de volum n:

$$\bar{x}_p = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}} \tag{3.11}$$

- pentru o populație de volum N:

$$\bar{\mu}_p = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}} \tag{3.12}$$

Între cele trei mărimi medii cu aplicații speciale prezentate anterior și media aritmetică există următoarea relație:

$$\bar{x}_h \leq \bar{x}_g \leq \bar{x} \leq \bar{x}_p \tag{3.13}$$

Egalitatea între aceste mărimi medii are loc doar în situația când toate variantele variabilei statistice X sunt egale între ele: $x_1 = x_2 = \cdots = x_n = constant$.

3.1.3. MEDIANA

Mediana sau valoarea mijlocie reprezintă acea valoare a unei variabile statistice care se situează în centrul seriei ordonate crescător sau descrescător, având proprietatea că numărul variantelor mai mari decât ea este egal cu numărul variantelor mai mici.

Scanned with OKEN Scanner

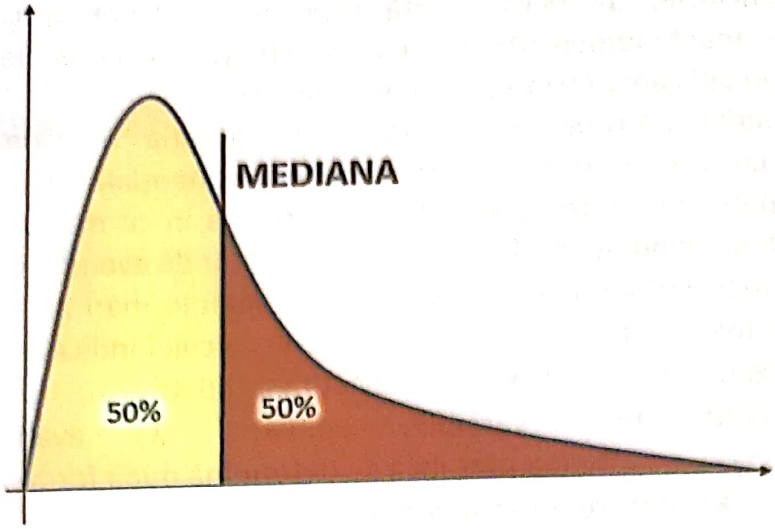


Figura 3.1. Mediana unei variabile statistice

Mediana poate fi definită ca fiind egală cu varianta x_i a variabilei statistice X, pentru care perpendiculara ridicată pe axa absciselor din punctul x_i împarte suprafața mărginită de curba frecvențelor în două părți egale (figura 3.1).

Este un indicator robust al tendinței centrale, mai puțin influențat de valorile extreme comparativ cu media aritmetică și mai stabil la fluctuațiile de selecție. Are o bază de calcul mai extinsă față media aritmetică, putând fi determinată și în cazul variabilelor categoriale, de tip ordinal.

Fie o serie statistică simplă, ordonată de forma:

$$x_1 \leq x_2 \leq \ldots \leq x_k \leq x_{k+1} \leq x_{k+2} \leq \ldots \leq x_n.$$

În determinarea efectivă a medianei ne putem situa în una dintre următoarele situații:

- Serii cu un *număr impar de termeni*: $n = 2k + 1$

Atunci locul medianei sau numărul de ordine al nivelului variabilei egal cu mediana va fi:

$$U^{me} = \frac{n+1}{2} = \frac{(2k+1)+1}{2} = k+1 \quad \Rightarrow \quad Me = x_{k+1} \tag{3.14}$$

- Serii cu un *număr par de termeni*: $n = 2k$

$$U^{me} = \frac{n+1}{2} = \frac{2k+1}{2} = k + \frac{1}{2} \quad \Rightarrow \quad Me = \frac{x_k + x_{k+1}}{2} \tag{3.15}$$

În acest caz, mediana se găsește în mijlocul intervalului dintre variantele x_p *și* x_{p+1}, (zona mediană).

Exemplul 1: Presupunem că avem vârstele unui număr de 5 copii (număr impar de termeni) ordonate crescător: 5, 7, 10, 13, 15. Mediana seriei va fi:

Scanned with OKEN Scanner

$$Me = 10\ ani$$

Exemplul 2: Considerăm că avem vârstele unui număr de 10 pacienți (număr par de termeni) ordonate crescător: 35, 43, 47, 56, 56, 58, 62, 63, 66, 68. Mediana seriei (vârsta mediană) va fi:

$$Me = \frac{56 + 58}{2} = 57\ ani$$

3.1.4. MODUL

Modul este o mărime medie de poziție, care se poate determina doar în cazul seriilor cu frecvență, și care reprezintă valoarea variabilei statistice care se întâlnește cel mai des (figura 3.2). Valoarea opusă modului, întâlnită în cazul distribuțiilor în formă de J și U, se numește antimod (valoarea cea mai rar întâlnită).

Modul este un indicator corespunzător al tendinței centrale numai în cazul distribuțiilor care se apropie de cea normală, cu un singur punct de maxim (unimodale). Acesta poate fi utilizat și în cazul centralizării variantelor variabilelor statistice categoriale, atât de tip nominal cât și ordinal. De exemplu, la gruparea pacienților în funcție de grupa de sânge din capitolul anterior (figura 2.5), observăm faptul că valoarea modală (cu frecvența maximă de apariție) este înregistrată pentru grupa B (36 de pacienți).

În practică, în cazul variabilelor statistice cantitative de tip continuu, pentru determinarea modului, deosebim două situații distincte:

a) *Serii cu frecvență pe variante ale caracteristicii*

Modul este egal cu varianta caracteristicii care are frecvență maximă: dacă considerăm $n_i = n_{max} \Rightarrow Mo = x_i$. În acest caz, modul este bine determinat și este egal cu acea variantă a variabilei care rezultă imediat din citirea tabelului în care este prezentată seria statistică.

b) *Serii cu frecvență pe intervale de variație a caracteristicii*

Se determină intervalul cu frecvență maximă sau intervalul modal (x_{i-1}, x_i):

În cadrul intervalului modal, valoarea căutată a modului va fi:

$$Mo = x_{i-1} + (x_i - x_{i-1}) \frac{(n_i - n_{i-1})}{(n_i - n_{i-1}) + (n_i - n_{i+1})} \qquad (3.16)$$

unde: x_{i-1} și x_i sunt limitele inferioară și superioară ale intervalului modal; n_i = frecventa intervalului modal, n_{i-1} = frecventa intervalului anterior și n_{i+1}= frecventa intervalului următor.

Scanned with OKEN Scanner

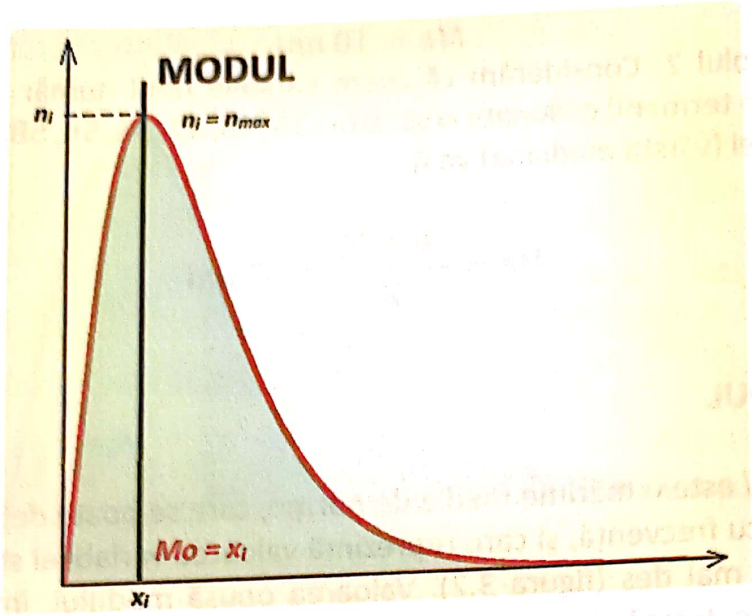


Figura 3.2. Modul în cazul unei variabile statistice continue

Exemplu: Presupunem că avem o distribuție statistică a variabilei înălțime pentru un eșantion format din 60 de persoane de sex masculin, grupat pe intervale egale de mărime 10 cm. Intervalul cu frecvență maximă (23 de persoane din 56) este 1,80 – 1,90 și acesta va conține valoarea căutată a înălțimii modale a grupului (tabelul 3.3).

Tabelul 3.3. Distribuția variabilei „înălțime" pe intervale de variație

Înălțimea (*m*)	**Nr. persoane de sex masculin**
1,40 - 1,50	1
1,50 - 1,60	3
1,60 - 1,70	8
1,70 - 1,80	10
1,80 - 1,90	**23**
1,90 – 2,00	15
TOTAL	**60**

Pentru a determina valoarea înălțimii celor mai mulți indivizi vom aplica formula de calcul 3.16 pentru situația în care variabila este distribuită pe intervale:

$$Mo = 1{,}80 + (1{,}90 - 1{,}80) \cdot \frac{(23 - 10)}{(23 - 10) + (23 - 15)} = 1{,}86\, m$$

Scanned with OKEN Scanner

În concluzie, valoarea modală a înălțimii persoanelor din eșantionul analizat este de 1,86 m (înălțimea cea mai frecventă în cadrul lotului).

3.2. INDICATORII VARIAȚIEI SAU DISPERSIEI

În situația în care ținem cont doar de mărimile medii într-un set de date sau realizăm o comparație între diferitele distribuții statistice utilizând exclusiv valorile tendinței centrale putem ajunge deseori în situația de a trage concluzii eronate. În completarea informațiilor furnizate valorile centrale ale unei serii statistice trebuie să avem în vedere dispersia, variabilitatea sau împrăștierea variantelor variabilelor.

Indicatorii variației măsoară *influența factorilor neesențiali,* care determină împrăștierea nivelurilor individuale ale unei variabile în jurul nivelului ei mediu. Se utilizează pentru caracterizarea omogenității colectivităților (o împrăștiere mai mică determină un grad mai mare de omogenitate și invers).

Gradul de împrăștiere exprimă și măsura în care o medie este reprezentativă pentru o anumită distribuție statistică (o valoare mică a indicatorilor variației înseamnă o concentrare ridicată a valorilor unei variabile în jurul tendinței ei centrale, ceea ce conferă mediei un caracter reprezentativ și invers).

Variația nivelurilor individuale ale unei caracteristici se poate măsura prin abaterea acestora (diferența) față de o mărime medie oarecare (se folosește cel mai frecvent media aritmetică).

Indicatorii variației exprimă fie distanța dintre două niveluri oarecare ale unei variabile, fie distanța medie dintre toate nivelurile înregistrate, considerate două câte două, fie distanța medie dintre nivelurile caracteristicii tuturor unităților și mărimea lor medie.

Indicatorii variației (dispersiei) sunt de două feluri: simpli și sintetici.

- *Indicatorii simpli* ai variației măsoară abaterea fiecărei variante a unei variabile față de nivelul ei mediu sau față de un alt nivel cu o anumită semnificație;
- *Indicatorii sintetici* ai variației măsoară împrăștierea medie a tuturor variantelor unei variabile în jurul nivelului ei mediu. Ei caracterizează complet gradul de împrăștiere a variantelor unei variabile în cadrul distribuțiilor statistice.

Scanned with OKEN Scanner

3.2.1. AMPLITUDINEA VARIAȚIEI

Amplitudinea variației oferă o imagine generală asupra împrăștierii datelor și reprezintă distanța dintre nivelul maxim și nivelul minim al variabilei. Acest indicator simplu al variației se poate determina astfel:

$$A = x_{maxim} - x_{minim} \tag{3.17}$$

În cazul grupării datelor pe intervale, se determină amplitudinea clasei ca diferență între valorile extreme ale fiecărui interval. Cu cât este mai mică valoarea acestui indicator, cu atât lotul este mai omogen.

Amplitudinea variației prezintă o serie de deficiențe:

- se bazează doar pe două valori (nivelul minim și cel maxim), neluând în considerare celelalte niveluri ale variabilei studiate;
- depinde de eșantionul de lucru particular, înregistrând variații pentru fiecare eșantion în parte ;
- nu ține cont de tipul de repartiție al variabilei statistice studiate.

Exemplu: utilizând distribuția variabilei înălțime prezentată anterior în tabelul 3.1, putem determina amplitudinea variației astfel:

$$A = 187 - 166 = 21\ cm$$

Putem în consecință formula o primă concluzie asupra nivelului variabilității din eșantionul de 10 persoane de sex masculin studiat, respectiv faptul că plaja de variație a înălțimii tuturor indivizilor din lot este de 21 cm, fiind cuprinsă între un nivel minim de 166 cm și un nivel maxim de 187 cm.

3.2.2. VARIANȚA ȘI DEVIAȚIA STANDARD

Varianța înlătură deficiența amplitudinii variației, utilizând în cadrul formulei toate variantele variabilei studiate, nu doar valorile minimă și maximă, măsurând în acest fel nivelul mediu al împrăștierii tuturor variantelor variabilei studiate de la media lor aritmetică.

Varianța sau dispersia se calculează ca o medie aritmetică a pătratelor abaterilor nivelurilor individuale ale unei variabile de la nivelul lor mediu.

- pentru un eșantion de volum n:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \tag{3.18}$$

Scanned with OKEN Scanner

- pentru o populație de volum N:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} \tag{3.19}$$

Tabelul 3.4. Determinarea variației pentru variabila „înălțime"

Persoana	Înălțimea x_i	Diferențele $x_i - \overline{x}$	Diferențele $(x_i - \overline{x})^2$
1	170	-6	36
2	177	1	1
3	175	-1	1
4	185	9	81
5	182	6	36
6	168	-8	64
7	172	-4	16
8	178	2	4
9	166	-10	100
10	187	11	121
Total	**1760**	**0**	**460**

Vom determina pentru început media variabilei „înălțime", utilizând formula 3.1:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1760}{10} = 176\ cm$$

Ulterior, calculăm în tabelul 3.4 diferențele dintre fiecare nivel individual al înălțimii și media seriei, pe care le ridicăm la pătrat și le însumăm în ultima coloană. Aplicăm în final formula 3.18 a variației pentru eșantionul analizat:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{460}{10-1} = 51{,}11\ cm^2$$

Deși prezintă o mare importanță în statistică, varianța nu are o interpretare concretă în practică, deoarece ia în calcul pătratele abaterilor variantelor unei variabile față de media lor. În consecință, unitatea de măsură pentru valoarea finală a indicatorului va fi și ea ridicată la puterea a doua. Mărimea variației este proporțională cu gradul de variabilitate dintr-o distribuție sau cu gradul ei de împrăștiere.

Scanned with OKEN Scanner

Deviația standard se determină prin extragerea rădăcinii pătrate din dispersie, reprezentând media pătratică a abaterilor nivelurilor individuale ale caracteristicii față de media lor.

- pentru un eșantion de volum n:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{3.20}$$

- pentru o populație de volum N:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}} \tag{3.21}$$

Continuând exemplul anterior, aplicăm formula 3.20 pentru determinarea deviației standard a eșantionului pentru variabila „înălțime":

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{s^2} = \sqrt{51{,}11} = 7{,}15\ cm$$

Putem concluziona astfel că, în medie indivizii din eșantionul cercetat se abat în medie de la media aritmetică a înălțimii grupului de 176 cm cu ± 7,15 cm.

Deviația standard se exprimă în aceleași unități de măsură ca și variabila luată în calcul, neputând însă fi folosită pentru caracterizarea împrăștierii unor colectivități care au caracteristicile exprimate în unități de măsură diferite. De exemplu, dacă dorim să realizăm o comparație între deviațiile standard calculate pe baza a două eșantioane în care valorile variabilei „înălțime" sunt exprimate în unități de măsură diferite (metri și picioare), nu am putea utiliza în mod direct deviația sau abaterea standard.

3.2.3. COEFICIENTUL DE VARIAȚIE

Coeficientul de variație este o măsură relativă a dispersării datelor față de medie și se calculează ca raport între deviația standard și nivelul mediu al unei variabile. Acesta se exprimă în procente, făcând abstracție de unitățile de măsură concrete ale variabilelor. În consecință coeficientul de variație poate fi utilizat în comparația gradului de împrăștiere a distribuțiilor, care sunt exprimate în unități de măsură diferite.

Scanned with OKEN Scanner

- pentru un eșantion de volum n:

$$v_s = \frac{s}{\bar{x}} \cdot 100 \quad (3.22)$$

- pentru o populație de volum N:

$$v_\sigma = \frac{\sigma}{\mu} \cdot 100 \quad (3.23)$$

Revenind la datele din tabelul 3.4, vom aplica formula 3.22 pentru a obține valoarea coeficientului de variație exprimată procentual:

$$v_s = \frac{s}{\bar{x}} \cdot 100 = \frac{7{,}15}{176} \cdot 100 = 4{,}06\%$$

În acest moment putem vorbi despre un nivel mediu de variabilitate de aproximativ 4% în eșantionul format studiat, sau cu alte cuvinte înălțimile celor 10 persoane se abat în medie de la înălțimea medie a grupului de 176 cm cu aproximativ ± 4%. În acest mod a fost depășită limitarea impusă de exprimarea în valoare absolută a deviației standard și putem realiza comparații între diferite loturi, chiar și în situația utilizării unor unități de măsură diferite.

3.2.4. CUARTILELE, DECILELE ȘI CENTILELE

În plus față de acești indicatori ai împrăștierii există și alte metode pentru a studia nivelul variabilității unui set de date. Una dintre acestea constă în determinarea poziției valorilor care împart un set de observații în 4, 10 sau 100 de părți egale (cuartilele, decilele și centilele).

Alături de mediană, există și alte mărimi de poziție care împart numărul variantelor unei variabile statistice, ordonate crescător sau descrescător, într-un număr de părți egale cu n / k. Aceste mărimi sunt cunoscute sub denumirea de *cuantile de ordinul* k. Ele se determină în mod asemănător cu mediana.

Dacă: $k = 2$ cuantila poartă denumirea de *mediană*;
$k = 4$ cuantilele poartă denumirea de *cuartile* (Q_1, Q_2, Q_3);
$k = 10$ cuantilele se numesc *decile* $(D_1, D_2, D_3, \dots, D_9)$;
$k = 100$ cuantilele se numesc *centile* $(C_1, C_2, C_3, \dots, C_{99})$.

Astfel, există trei *cuartile* care împart un șir de valori ale unei variabile ordonate crescător în patru părți, în așa fel încât fiecare parte să cuprindă același număr de elemente. Prima cuartilă (Q_1) este acea valoare a unei variabile statistice care împarte o distribuție în două părți, astfel încât un sfert din

Scanned with OKEN Scanner

variantele acesteia au valori mai mici decât Q_1 , iar trei sferturi au valori mai mari decât Q_1. Cuartila a doua (Q_2) împarte numărul variantelor unei variabile statistice în două părți egale, astfel încât jumătate dintre variante au valori mai mici decât Q_2 și jumătate au valori mai mari decât Q_2. Q_2 coincide, deci, cu mediana. Cuartila a treia (Q_3) împarte numărul termenilor unei serii în două părți, astfel încât trei sferturi dintre ei au valori mai mici decât Q_3 și un sfert au valori mai mari decât Q_3 (figura 3.3)

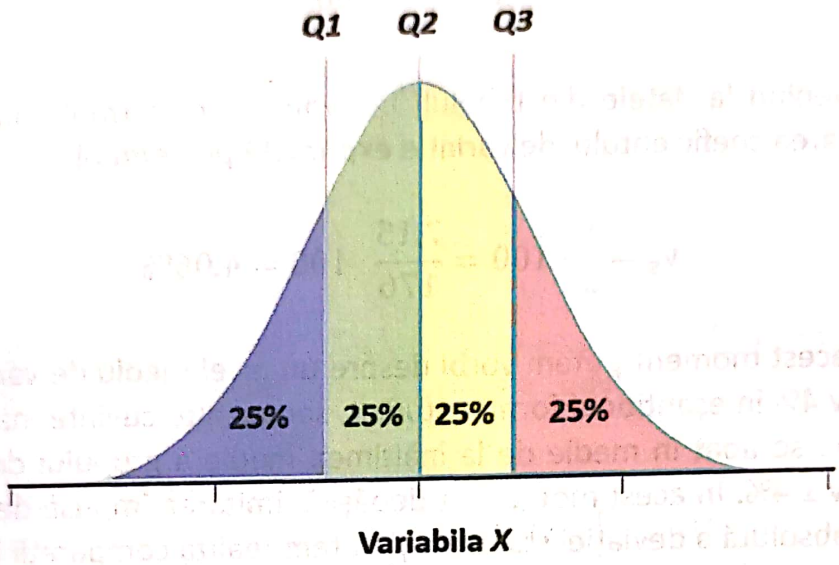


Figura 3.3. Poziția cuartilelor în cadrul distribuției unei variabile

Decilele, în număr de 9, împart un șir de valori ale unei variabile ordonate în 10 părți (intervale) egale. *Centilele*, în număr de 99, împart variantele variabilei ordonate în 100 de părți egale.

Locația sau poziția unei cuantile k în cadrul șirului ordonat al variantelor variabilei $X(x_1, x_2, x_3, \dots, x_n)$ se determină după relația:

$$L_k = (n+1) \cdot \frac{k}{100} \tag{3.24}$$

unde n reprezintă numărul total de variante x_i ale variabilei X.

Astfel, locul celor 3 cuartile în cadrul distribuției variabilei X va fi:

$$L_{Q_1} = L_{25} = (n+1) \cdot \frac{25}{100} = \frac{n+1}{4}$$

$$L_{Q_2} = L_{Me} = L_{50} = (n+1) \cdot \frac{50}{100} = \frac{n+1}{2}$$

$$L_{Q_3} = L_{75} = (n+1) \cdot \frac{75}{100} = \frac{3 \cdot (n+1)}{4}$$

Scanned with OKEN Scanner

Un indicator similar amplitudinii variației, care ne oferă o imagine preliminară asupra nivelului variabilității într-un set de date, este reprezentat de *intervalul intercuartilic*, care reflectă variabilitatea existentă în cadrul celor 50% dintre variante situate în centrul distribuției. Poate fi privit ca un indicator al variabilității observațiilor relevante din cadrul seriei studiate. Un interval intercuartilic larg este un indicator al unui nivel ridicat de împrăștiere în cadrul intervalului care cuprinde 50% dintre observațiile relevante din centrul seriei statistice. Se poate determina astfel:

$$I_Q = Q_3 - Q_1 \tag{3.25}$$

Cu ajutorul cuartilelor poate fi construit un grafic deosebit de util, care poate caracteriza forma generală a distribuției unei variabile statistice, evidențiind în același timp și gradul de simetrie al acesteia cât și prezența potențialelor valori extreme. Valorile neobișnuite sau extreme (foarte depărtate de celelalte variante ale variabilei, în ambele sensuri) se numesc outlieri și, în mod convențional, se situează în afara în afara intervalului:

$$[Q_1 - 1{,}5 \cdot (Q_3 - Q_1),\ Q_3 + 1{,}5 \cdot (Q_3 - Q_1)] \tag{3.26}$$

Graficul „Box-plot” se bazează pe cinci indicatori statistici descriptivi, respectiv valorile minimă și maximă, prima și a treia cuartilă și valoarea mediană a seriei (figura 3.4).

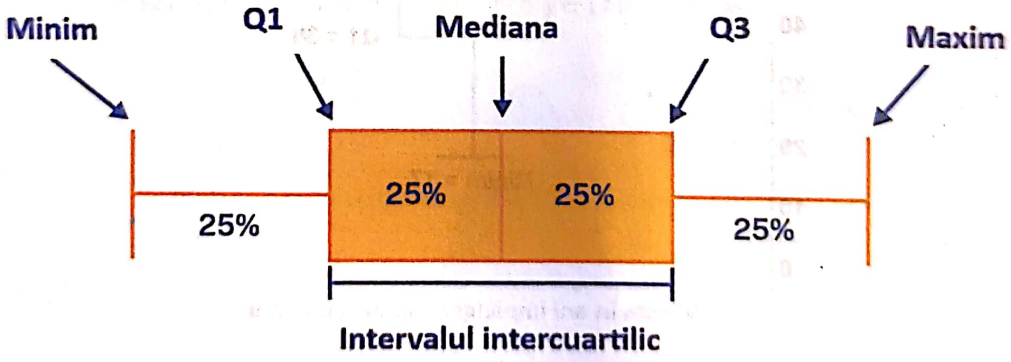


Figura 3.4. Diagrama „Box-plot”

În vederea exemplificării, revenim la tabelul 2.2 prezentat în capitolul anterior, în care am ordonat cei 100 de pacienți din eșantion crescător în funcție de variantele variabilei „vârstă”, de la un nivel minim de 17 ani până la valoarea maximă de 75 de ani. Determinăm locul celor 3 cuartile în cadrul distribuției variabilei vârstă:

Scanned with OKEN Scanner

$$L_{Q_1} = \frac{n+1}{4} = \frac{100+1}{4} = 25{,}25$$
$$L_{Q_2} = L_{Me} = \frac{n+1}{2} = \frac{100+1}{2} = 50{,}50$$
$$L_{Q_3} = \frac{3\cdot(n+1)}{4} = \frac{3\cdot(100+1)}{4} = 75{,}75$$

Rezultă că, cuartila Q_1 va fi situată, în cadrul seriei ordonate crescător, între temenii 25 (valoarea 36) și 26 (valoarea 37), la distanță de 0,25 de primul (36) și 0,75 de al doilea (37). Distanța dintre valorile 36 și 37 este de mărime 1, astfel încât aplicăm coeficientul de 0,25 la această valoare: $0{,}25 \cdot 1 = 0{,}25$. Adunăm 0,25 la 36 (termenul din stânga) și obținem astfel valoarea primei cuartile: $Q_1 = 36{,}25\ ani$. În mod similar, determinăm cuartila $Q_3 = 56{,}75$. Mediana va fi situată între termenii 50 și 51, la distanță egală față de ambii. Dar $x_{50} = x_{51} = 48\ ani$, iar mediana, calculată ca medie aritmetică simplă a acestora va avea aceeași valoare: $Me = 48\ ani$.

Pornind de la aceste valori, construim diagrama Box-plot a vârstelor celor 100 de pacienți (figura 3.5).

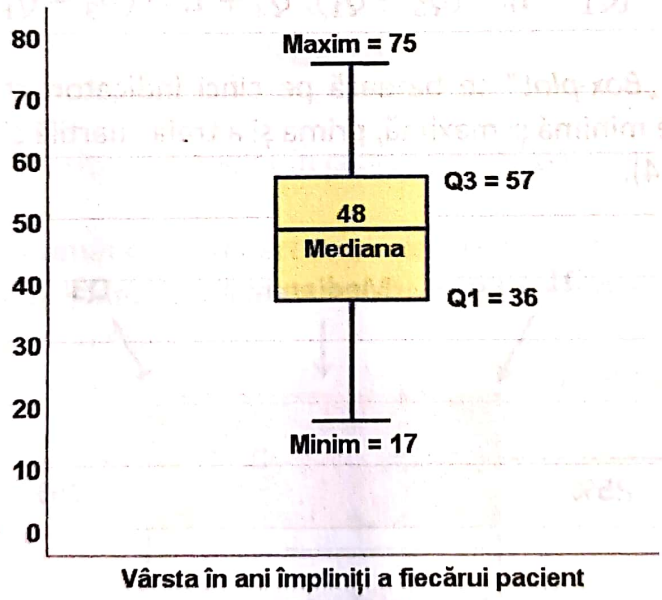


Figura 3.5. Diagrama „Box-plot" a vârstelor celor 100 de pacienți

Analizând graficul, rezultă imediat faptul că vârsta mediană a pacienților a fost de 48 de ani, 25% dintre aceștia au sub 36 de ani, iar 25% peste 57 de ani. Jumătate dintre pacienți au vârstele cuprinse între 36 și 57 de ani. Distribuția vârstelor este aproximativ simetrică, distanța dintre limita inferioară și Q_1 este relativ similară cu cea dintre limita superioară și Q_3. De asemenea, și aria din dreptunghi cuprinsă între Me și Q_1 este aproape egală cu aria dintre Me și Q_3.

Scanned with OKEN Scanner

Nu este semnalată în cadrul graficului existența unei valori anormale, extreme, foarte mici sau foarte mari în comparație cu celelalte valori ale variabilei vârstă, care să fie neobișnuit de departe de restul valorilor seriei (vârsta niciunui pacient din eșantion nu se situează în afara limitelor intervalului prezentat în formula 3.26):

$$[36 - 1{,}5 \cdot (57 - 36);\ 57 + 1{,}5 \cdot (57 - 36)] = [5{,}50;\ 87{,}50]$$

În concluzie, distribuția vârstelor pacienților este aproximativ simetrică față de valoarea medianei, nefiind evidențiată prezența niciunui outlier. Cele mai relevante 50% dintre valorile variabilei „vârstă", localizate în zona centrală a distribuției, între cuartila 1 (36 ani) si cuartila 3 (57 ani), sunt poziționate într-un interval de doar 21 de ani, comparativ cu întreaga plajă de valori a variabilei care se extinde pe 58 de ani.

3.3. INDICATORII FORMEI DISTRIBUȚIILOR

Pe lângă tendința centrală și dispersia valorilor unei variabile, o altă caracteristică importantă a unei distribuții statistice este reprezentată de forma acesteia. Un caz particular, având implicații teoretice și practice deosebite în statistică, se referă la situația când media, mediana și modul sunt egale. Este vorba despre bine-cunoscuta *curbă normală*, care are alura unui clopot. Formula matematică a acestei distribuții statistice a fost inițial publicată către matematicianul francez Abraham de Moivre (1667-1754).

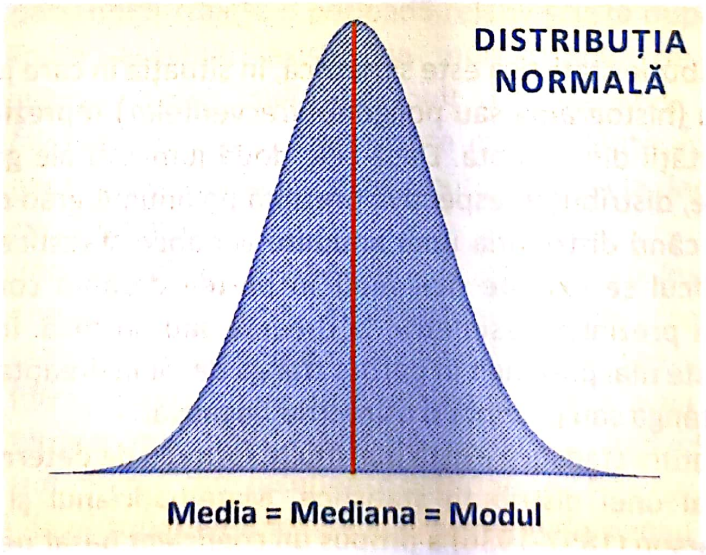


Figura 3.6. Distribuția normală

Scanned with OKEN Scanner

Contribuții importante a avut și matematicianul german Carl Friederich Gauss (1777-1855), din acest motiv distribuția normală purtând și numele de distribuție gausiană.

Distribuția normală este una dintre cele mai importante distribuții de probabilitate din statistică deoarece descrie cu exactitate distribuția valorilor pentru multe fenomene naturale: greutatea, înălțimea, scorurile IQ, cantitatea de precipitații, notele la un examen standardizat, erorile de măsurare, etc.

Astfel, greutatea la naștere a nou-născuților urmează o distribuție aproximativ normală, majoritatea dintre aceștia având o greutate apropiată de medie, numărul lor (frecvența lor de apariție) scăzând drastic pe măsură ce ne apropiem de valorile extreme. Un alt exemplu ar putea fi reprezentat de presiunea arterială sau nivelul hemoglobinei în cazul unor populații sănătoase, situații în care marea majoritate a valorilor acestor variabile sunt grupate în jurul mediilor lor. De asemenea, timpul necesar pentru recuperare în urma unei intervenții chirurgicale sau erorile de măsurare în cadrul unui experiment sunt alte exemple de variabile statistice care urmează cu aproximație o lege normală de repartiție.

Distribuția normală are mai multe proprietăți, dintre care enumerăm:

- are o repartiție *simetrică* în jurul mediei aritmetice
- media aritmetică, mediana și modul coincid
- suprafața delimitată de curbă și axa orizontală este egală cu 1
- este complet determinată de 2 parametri (media aritmetică și deviația standard)

3.3.1. COEFICIENTUL DE ASIMETRIE

O distribuție statistică este simetrică, în situația în care jumătatea stângă a graficului său (histograma sau poligonul frecvențelor) reprezintă o imagine în oglindă a jumătății din dreapta. Dacă cele două jumătăți ale graficului nu sunt perfect identice, distribuția respectivă prezintă un anumit grad de asimetrie.

Atunci când distribuția unei anumite variabile statistice este asimetrică pentru că graficul se extinde mai mult în partea dreaptă comparativ cu cea stângă, aceasta prezintă o asimetrie la dreapta sau pozitivă. În situația opusă, când graficul este mai prelungit în partea stângă decât în dreapta, distribuția este asimetrică la stânga sau prezintă o asimetrie negativă.

În literatura statistică există mai multe formule de determinare a gradului de asimetrie al unei distribuții statistice. Matematicianul și biostatisticianul englez Karl Pearson (1857-1936) a propus un coeficient bazat pe diferența dintre media aritmetică ($\bar{x}$) și mod (Mo), raportată la deviația standard a variabilei (σ):

Scanned with OKEN Scanner

$$S_k = \frac{\bar{x} - Mo}{\sigma} \approx \frac{3 \cdot (\bar{x} - Me)}{\sigma} \quad (3.27)$$

Coeficientul de asimetrie al lui Pearson poate lua valori între -3 și 3, valoarea zero fiind specifică unei distribuții simetrice, în care media aritmetică, mediana și modul coincid. De exemplu, o valoare de -2,6 indică o asimetrie negativă (la stânga) importantă, iar o valoare de +1,5 arată o asimetrie pozitivă la dreapta (pozitivă) moderată.

Din punct de vedere grafic, în cazul caracterizării gradului de asimetrie al unei distribuții statistice, ne putem situa într-una dintre situațiile prezentate în figura 3.7.

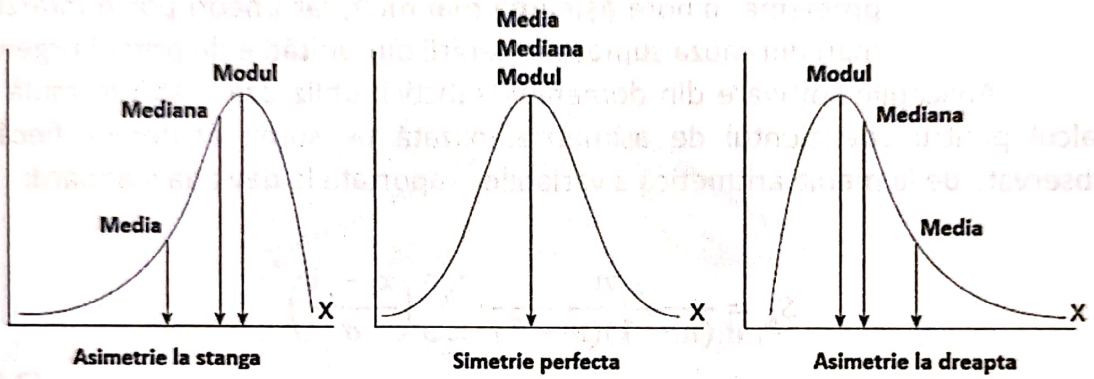


Figura 3.7. Tipuri de asimetrie a unei distribuții statistice

Exemple de *distribuții asimetrice la stânga* în domeniul medical:

- *Vârsta la deces în cazul unor boli incurabile*: majoritatea pacienților trăiesc o perioadă relativ scurtă după diagnostic. Cu toate acestea, există un număr mic de pacienți care supraviețuiesc mult mai mult datorită unor factori genetici favorabili sau unor tratamente mai eficiente. În acest fel se va genera o coadă lungă a distribuției vârstei la deces a pacienților spre stânga;
- *Rata de filtrare glomerulară la pacienții cu insuficiență renală cronică*: în cazul insuficienței renale cronice avansate, majoritatea pacienților au valori foarte scăzute ale ratei de filtrare glomerulare, însă unii pacienți pot avea aceasta rată de filtrare mai mare, datorită variațiilor individuale sau unor tratamente care încetinesc deteriorarea funcției renale.

Exemple de *distribuții asimetrice la dreapta* în domeniul medical:

- *Nivelul enzimelor TGO/ TGP în bolile hepatice*: în bolile hepatice ușoare sau moderate, enzimele hepatice sunt puțin crescute. Cu toate acestea, în cazurile severe (hepatită virală acută sau

Scanned with OKEN Scanner

insuficiență hepatică fulminantă), valorile acestor enzime pot crește extrem de mult, depășind de zeci sau sute de ori limita normală, formând astfel o coadă lungă a distribuției spre dreapta;

- *Nivelul glicemiei în diabet*: majoritatea persoanelor cu diabet controlat au glicemia cuprinsă între 100-200 mg/dL. Totuși, în cazurile necontrolate sau în episoadele de hiperglicemie severă, valorile glicemiei pot ajunge la 600-800 mg/dL, creând în acest mod o coadă lungă a distribuției spre dreapta;
- *Timpul de așteptare în Urgențe*: pacienții sunt triați în funcție de gravitate, astfel încât cazurile urgente sunt tratate imediat. Cei cu probleme minore așteaptă mai mult, iar uneori pot fi întârzieri mari din cauza supraaglomerării din Unitățile de primiri urgențe.

Aplicațiile software din domeniul statisticii utilizează o altă formulă de calcul pentru coeficientul de asimetrie, bazată pe suma abaterilor fiecărei observații de la media aritmetică a variabilei, raportate la deviația standard:

$$S_k = \frac{n}{(n-1)(n-2)} \cdot \sum_{i=1}^{k} \left(\frac{x_i - \bar{x}}{\sigma}\right)^3 \qquad (3.28)$$

unde:

n = numărul de variante ale variabilei studiate
x_i = varianta i a variabilei X
$\bar{x}$ = media variabilei X
σ = deviația standard a variabilei X

În funcție de valoarea coeficientului de asimetrie, deosebim 3 cazuri:

- $S_k = 0$ distribuție perfect simetrică;
- $S_k > 0$ distribuție asimetrică la dreapta;
- $S_k < 0$ distribuție asimetrică la stânga.

Distribuțiile asimetrice sunt des întâlnite în medicină și pot influența diagnosticul, tratamentul și interpretarea rezultatelor. Medicii trebuie să țină cont de aceste distribuții atunci când analizează datele, deoarece media aritmetică simplă poate genera rezultate nerelevante în astfel de cazuri. Prin urmare, în locul mediei, va fi utilizată în analiză mediana pentru a descrie mai precis centrul distribuției variabilei de interes.

3.3.2. COEFICIENTUL DE BOLTIRE

Boltirea reprezintă o măsură a gradului în care o distribuție a unei variabile statistice este mai înaltă sau mai aplatizată în comparație cu distribuția normală, al cărei grafic este de forma unui clopot (figura 3.8).

Scanned with OKEN Scanner

Dacă o variabilă statistică prezintă o pondere însemnată a variantelor sau observațiilor în cozile laterale ale graficului (histograma sau poligonul frecvențelor) și implicit un procent mai redus al lor la mijlocul seriei, suntem în situația unei distribuții platicurtice. Este o distribuție în care rezultatele sunt foarte împrăștiate fata de medie, indicând un grad ridicat de eterogenitate a nivelurilor individuale ale variabilei.

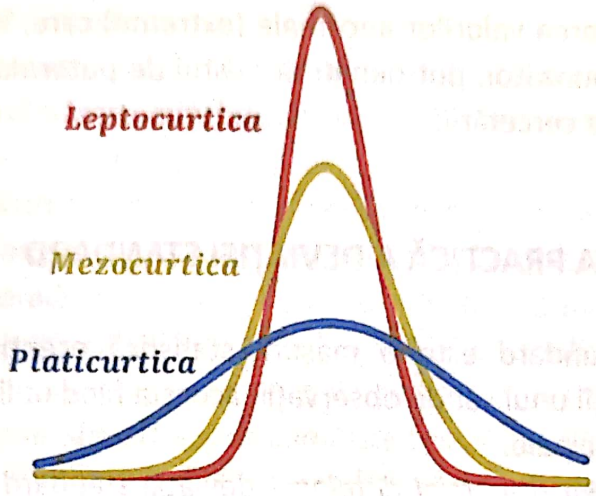


Figura 3.8. Tipuri de boltire a unei distribuții statistice

În situația în care, o variabilă prezintă o pondere redusă a variantelor sau observațiilor sale în cozile laterale ale graficului și implicit un procent mai însemnat al lor la mijlocul seriei, suntem în situația unei distribuții leptocurtice. O distribuție leptocurtică, ascuțită, arata ca datele sunt foarte grupate si apropiate de medie, lotul de subiecți având un mare grad de omogenitate a valorilor înregistrate.

Distribuția normală este o distribuție mezocurtică.

Pentru caracterizarea nivelului de boltire Karl Pearson a propus un coeficient de boltire:

$$\gamma_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n \cdot \sigma^4} - 3 \qquad (3.29)$$

unde:

n = numărul de variante ale variabilei studiate
x_i = varianta i a variabilei X
$\bar{x}$ = media variabilei X
σ = deviația standard a variabilei X

În funcție de valoarea coeficientului de boltire, deosebim 3 cazuri:

- $\gamma_2 = 0$ distribuție perfect simetrică (mezocurtică);
- $\gamma_2 > 0$ distribuție leptocurtică;
- $\gamma_2 < 0$ distribuție platicurtică.

Scanned with OKEN Scanner

Indicatorii statistici descriptivi sunt esențiali pentru înțelegerea și interpretarea datelor. Ei oferă informații importante despre tendința centrală de dezvoltare a unui fenomen, nivelul împrăștierii datelor variabilei studiate și forma funcțională a unei distribuții statistice. Aceștia permit într-o primă etapă conturarea unei imagini clare asupra tipului de distribuție a variabilelor statistice, evaluarea preliminară a normalității datelor analizate (condiție esențială pentru aplicarea testelor statistice parametrice de verificare a ipotezelor statistice) precum și identificarea valorilor anormale (extreme) care, în situația în care nu sunt tratate corespunzător, pot denatura destul de puternic, în anumite situații practice, rezultatele cercetării.

3.4. IMPORTANȚA PRACTICĂ A DEVIAȚIEI STANDARD

Deviația standard este o măsură statistică esențială de evaluare a nivelului variabilității unui set de observații, aceasta fiind utilizată în mod practic în două direcții principale:

- *evaluarea dispersiei datelor* – deviația standard arată cât de mult variază valorile unui set de date față de media acestuia. O deviație standard mică indică un grad ridicat de omogenitate, datele fiind grupate în jurul mediei lor, în timp ce o valoare ridicată sugerează un nivel de împrăștiere semnificativ în cadrul distribuției;
- *compararea variabilității între seturi de date* – permite compararea gradului de dispersie între două sau mai multe seturi de date, oferind o perspectivă mai detaliată decât simpla analiză a mediilor. De multe ori în practică nu este suficientă comparația realizată între mediile a două sau mai multe variabile, fiind foarte important să cunoaștem cât de mult variază valorile individuale față de media fiecărui set de observații.

3.4.1. SCORURILE Z

Utilizând media și deviația standard putem obține poziția relativă a oricărei valori x_i a variabile statistice $X(x_1, x_2, x_3, \dots, x_i, \dots, x_n)$ supuse analizei. Presupunând că am determinat pe baza datelor din eșantion media aritmetică ($\bar{x}$) și deviația standard (σ), vom avea pentru fiecare variantă x_i a variabilei X o valoare asociată z_i care poartă numele de *„scorul z al acesteia"*:

$$z_i = \frac{x_i - \bar{x}}{s} \tag{3.30}$$

Scanned with OKEN Scanner

Scorul z_i este denumit frecvent „valoarea standardizată a nivelului x_i" al variabilei X și poate fi interpretat ca reprezentând numărul de deviații standard care separă varianta x_i a variabilei X de nivelul mediu $\bar{x}$ al seriei statistice.

Scorul z_i transformă datele variabilei X într-o *scară standardizată*, astfel încât să putem compara valori din distribuții diferite. Un scor z_i pozitiv înseamnă că valoarea este *peste medie*, iar un scor z_i negativ înseamnă că este *sub nivelul mediu*.

Proprietățile scorului z_i:

a) *compararea valorilor din seturi de date diferite* (scorurile z_i permit comparația între valorile unor variabile care sunt exprimate în unități de măsură diferite);

b) *identificarea valorilor extreme* (în situația în care un anumit scor z_i este foarte mare, în afara intervalului [-3, 3], această valoare poate fi considerată un *outlier* sau o valoare neobișnuită, care este semnificativ diferită de restul variantelor variabilei studiate);

c) scorurile z_i sunt *esențiale în aplicarea testelor de verificare a ipotezelor statistice*, așa cum este *testul Z*, utilizat pentru a verifica dacă un eșantion diferă semnificativ de populația totală din care a fost extras (de exemplu, diagnosticarea unor afecțiuni prin compararea rezultatelor testelor cu valorile normale).

Vom exemplifica, utilizând datele privind distribuția variabilei statistice „înălțime" pentru un eșantion format din 10 pacienți, prezentate inițial în tabelul 3.1.

Tabelul 3.5. Determinarea scorurilor z_i pentru variabila "înălțime"

Persoana	Înălțimea (x_i)	$x_i - \bar{x}$	$z_i = \frac{x_i - \bar{x}}{s}$
1	170	-6	-0,84
2	177	1	0,14
3	175	-1	-0,14
4	*185*	*9*	*1,26*
5	182	6	0,84
6	168	-8	-1,12
7	172	-4	-0,56
8	178	2	0,28
9	*166*	*-10*	*-1,40*
10	187	11	1,54
SUMA	**1760**	**0**	**0**
MEDIA ($\bar{x}$)	**176**	**0**	**0**
DEV STD (s)	**7,15**	**-**	**1**

Scanned with OKEN Scanner

Analizând tabelul 3.5, se observă imediat faptul că persoana nr. 4 are valoarea $z_i = 1{,}26$, aspect care indică că înălțimea acesteia de 185 cm este mai mare decât înălțimea medie a eșantionului de 176 cm cu 1,26 deviații standard. La polul opus, persoana nr. 9 se situează cu 1,40 deviații standard sub nivelul mediu al înălțimii eșantionului.

Concluzionând, scorul z_i pentru o variantă a unei anumite variabile poate fi interpretat ca o măsură a poziției relative a acesteia în cadrul seriei. Astfel, variantele a două variabile statistice diferite care au aceeași valoare a scorului z_i, dețin aceeași poziție în cadrul celor 2 distribuții ale lor, în sensul că sunt la același număr de deviații standard de media fiecărei variabile.

3.4.2. TEOREMA LUI CEBÎȘEV ȘI REGULA EMPIRICĂ

Matematicianul rus Pafnuti Lvovici Cebîșev (1821-1894) a dezvoltat o teoremă care ne permite să determinăm proporția minimă a variantelor unei variabile statistice care se situează într-un anumit număr de deviații standard de la media aritmetică a seriei.

Teorema lui Cebîșev: în cazul oricărui set de observații (eșantion sau populație), indiferent de tipul distribuției, proporția valorilor variabilei statistice care sunt cuprinse în intervalul $(\bar{x} - k \cdot \sigma,\ \bar{x} + k \cdot \sigma)$ este de cel puțin:

$$1 - \frac{1}{k^2} \tag{3.31}$$

unde k reprezintă orice valoare mai mare decât 1.

De exemplu, pentru a determina ce procent dintre variantele unei variabile X, indiferent de tipul distribuției acesteia, se vor situa în intervalul $(\bar{x} - 2 \cdot \sigma,\ \bar{x} + 2 \cdot \sigma)$, adică la ± 2 deviații standard în jurul mediei aritmetice, vom aplica teorema lui Cebîșev astfel:

$$1 - \frac{1}{k^2} = 1 - \frac{1}{2^2} = 1 - \frac{1}{4} = 0{,}75$$

În concluzie, cel puțin 0,75 sau 75% dintre observațiile unui set de date se vor situa în intervalul $(\bar{x} - 2 \cdot \sigma,\ \bar{x} + 2 \cdot \sigma)$.

Similar pentru intervalul $(\bar{x} - 3 \cdot \sigma,\ \bar{x} + 3 \cdot \sigma)$, vom obține o valoare de cel puțin 0,889 sau 88,9% dintre variantele variabilei.

Unul dintre avantajele teoremei lui Cebîșev este faptul că nu ține cont de tipul distribuției variabilei studiate. În cazul unei variabile distribuite aproximativ normal, dispersia sau împrăștierea variantelor seriei în jurul mediei aritmetice a acesteia poate fi realizată însă mult mai precis.

Scanned with OKEN Scanner

Regula empirică: pentru o distribuție statistică de frecvențe simetrică având alura unui clopot (figura 3.9), 68,27% dintre observații vor fi cuprinse între medie și plus minus o deviație standard, 95,45% dintre observații se vor situa între medie și plus minus două deviații standard și aproape toate observațiile (99,73%) se vor situa în intervalul $(\mu - 3 \cdot \sigma,\ \mu + 3 \cdot \sigma)$.

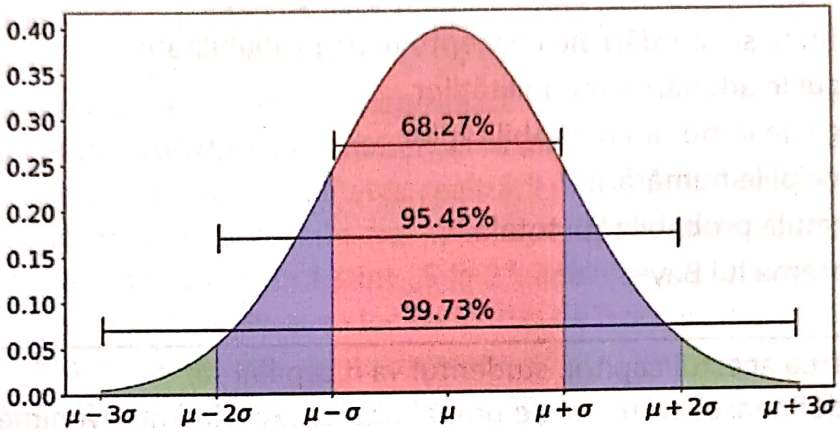


Figura 3.9. Regula empirică

În analiza practică se utilizează de multe ori intervalul în care există 95% șanse să se situeze anumită variantă a unei variabile statistice normal distribuite. Acest interval este de forma: $(\mu - 1{,}96 \cdot \sigma,\ \mu + 1{,}96 \cdot \sigma)$.

Scanned with OKEN Scanner

Capitolul. 4

CONCEPTE DE BAZĂ ALE TEORIEI PROBABILITĂȚILOR

SUBIECTE:

4.1. Definiții și abordări ale conceptului de probabilitate
4.2. Regulile adunării probabilităților
4.3. Regulile înmulțirii probabilităților
4.4. Principiile numărării
4.5. Formula probabilității totale
4.6. Teorema lui Bayes

La finalizarea acestui capitol, studentul va fi capabil să:
O4-1: să definească termenii de probabilitate, experiment, eveniment și rezultat
O4-2: să exemplifice noțiunile de probabilitate clasică, empirică și subiectivă
O4-3: să calculeze probabilități utilizând regulile adunării probabilităților
O4-4: să calculeze probabilități utilizând regulile înmulțirii probabilităților
O4-5: să calculeze numărul de rezultate utilizând principiile numărării

4.1. DEFINIȚII ȘI ABORDĂRI ALE CONCEPTULUI DE PROBABILITATE

Teoria probabilităților se ocupă cu studiul experimentelor cu rezultat aleatoriu (întâmplător) aceasta dezvoltându-se accelerat la începutului secolului al XVII-lea, fiind puternic impulsionată de extinderea jocurilor de noroc.

Așa cum am văzut în capitolele anterioare *statistica descriptivă* (histograma, poligonul frecvențelor, graficele de structură, diagramele Box Plot, indicatorii tendinței centrale, ai variației și formei distribuțiilor) are la bază analiza datelor colectate de la evenimente care au avut loc anterior. În următoarele capitole, accentul va fi pus cu precădere asupra șansei ca un anumit eveniment să se producă în viitor – aspect studiat de statistica inferențială.

Statistica inferențială urmărește să generalizeze la scara întregii populații concluziile obținute în urma analizei datelor la nivelul unui eșantion reprezentativ extras din aceasta. Deoarece acest demers implică un anumit nivel de incertitudine, în procesul complex de luare a deciziilor este importantă realizarea unei evaluări științifice a tuturor riscurilor care sunt cunoscute.

Pentru a realiza în mod corespunzător acest proces de evaluare este necesar să facem apel la teoria probabilităților sau „știința incertitudinii". Teoria

Scanned with OKEN Scanner

probabilităților dă posibilitatea tuturor factorilor de decizie implicați să analizeze riscurile și să minimizeze șansa de a accepta de exemplu o achiziție de echipamente electronice medicale care au anumite deficiențe.

Probabilitatea reprezintă o valoare cuprinsă între 0 și 1, care descrie posibilitatea relativă (șansa) ca un eveniment să aibă loc.

Probabilitatea se exprimă frecvent ca un număr zecimal (0,25 sau 0,50) sau procentual (25% sau 50%). O probabilitate egală cu 1 reprezintă ceva ce sigur se va întâmpla, iar o probabilitate egală cu zero ne arată ceva care nu se poate întâmpla (imposibil). Cu cât este mai apropiată de zero valoarea probabilității cu atât este mai puțin probabil ca evenimentul să aibă loc.

În anumite situații, posibilitatea realizării unui eveniment este exprimată prin utilizarea termenului *șansă*. De exemplu, o persoană afirmă că șansele ca un anumit eveniment să se producă sunt „5 la 2", adică, în situația în care avem 7 încercări (5+2) evenimentul va avea loc de 5 ori și de 2 ori nu se va realiza. Cu alte cuvinte, utilizând șansele ca un eveniment să se realizeze, putem calcula probabilitatea aferentă: 5 / (5+ 2) = 5 / 7 = 0,71 sau 71%.

Există o serie de noțiuni cheie în studierea probabilităților:

- *experimentul* (experiența) reprezintă realizarea practică a unui complex de condiții, corespunzătoare unui criteriu dat de cercetare a colectivităților statistice omogene;
- *proba* este realizarea (producerea) o singură dată a experienței;
- *evenimentul* este rezultatul concret al unei probe.

De exemplu: aruncarea unui zar perfect construit de un anumit număr de ori reprezintă un experiment, care are șase rezultate posibile (1,2,3,4,5,6), proba reprezintă aruncarea o singură dată a zarului, iar câteva posibile evenimente ar putea fi: observarea unui număr par sau observarea unui număr mai mare decât trei, etc.

Evenimentele pot fi clasificate în trei mari categorii:

- *sigure* (se produc obligatoriu la efectuarea unei probe);
- *imposibile* (nu se produc la efectuarea niciunei probe);
- *întâmplătoare* (poate sau nu să se producă la efectuarea unei singure probe).

De exemplu:

a) extragerea unei bile albe dintr-o urnă care conține numai bile albe (eveniment sigur)
b) apariția unei fețe cu 7 puncte la proba aruncării unui zar (eveniment imposibil)
c) apariția feței cu 5 puncte la aruncarea unui zar (eveniment întâmplător)

Evenimentele întâmplătoare (aleatorii) se supun unor legități statistice. Dacă revenim la proba aruncării o singură dată a unui zar, nu putem prevedea dacă vom obține fața cu numărul 5. Cu toate acestea, dacă vom realiza un număr

Scanned with OKEN Scanner

suficient de mare de probe în cadrul unui experiment, vom putea calcula cu un anumit grad de precizie numărul de apariții ale acestei fețe a zarului.

Evenimentele întâmplătoare pot fi *compatibile* (realizarea unuia nu exclude realizarea celuilalt extragerea unei cărți de treflă și a unui valet dintr-un pachet de cărți) și *incompatibile* (realizarea unuia exclude realizarea celuilalt – de exemplu extragerea unui as și extragerea unei dame dintr-un pachet de cărți).

Evenimentele pot fi *dependente* (realizarea unuia influențează probabilitatea realizării celuilalt) și *independente* (realizarea unuia nu influențează probabilitatea realizării celuilalt). Putem exemplifica astfel:

Evenimente independente:

1 – apariția feței 1 la aruncarea unui zar

2 – apariția feței 5 la o altă aruncare a zarului

Evenimente dependente:

1 – obținerea unui număr de 3 puncte la aruncarea a două zaruri

2 – apariția feței cu numărul 1 pe unul dintre aceste două zaruri

Probabilitatea unui eveniment poate fi definită în trei modalități principale: clasică, empirică și subiectivă. Primele două metode (clasică și empirică) sunt de natură obiectivă, fiind bazate pe informații și date reale, în timp ce probabilitatea subiectivă reprezintă o estimare personală privind posibilitatea producerii unui eveniment, bazată pe orice informație disponibilă în acel moment (este utilizată în situațiile existenței unui volum limitat de informații).

1. *Probabilitatea clasică* are la bază presupunerea că toate rezultatele unui experiment să fie egal posibile. Astfel, probabilitatea producerii unui anumit eveniment A se poate scrie sub forma:

$$P(A) = \frac{Numarul\ de\ rezultate\ favorabile\ producerii\ evenimentului\ A}{Numarul\ total\ de\ rezultate\ posibile\ ale\ experimentului} \quad (4.1)$$

De exemplu, probabilitatea de a obține numărul 6 la aruncarea unui zar este 1 / 6, iar probabilitatea de a obține un număr par la aruncarea aceluiași zar este de 3 / 6 = 1 / 2. La aruncarea simultană a două zaruri, probabilitatea de a obține dubla 6:6 sau suma cifrelor 12 este de 1 / 36.

Deoarece numărătorul probabilității producerii unui anumit eveniment A este întotdeauna mai mic sau egal cu numitorul acesteia rezultă că:

$$0 \leq P(A) \leq 1 \quad (4.2)$$

Astfel:

- *Probabilitatea evenimentului sigur este 1* (evenimentul are loc de fiecare dată la efectuarea experimentului)
- *Probabilitatea evenimentului imposibil este 0* (evenimentul nu se produce niciodată)

Scanned with OKEN Scanner

- *Probabilitatea unui eveniment aleatoriu (întâmplător) este cuprinsă între 0 și 1*

2. *Probabilitatea empirică* ca un anumit eveniment A să aibă loc în prezent se calculează în funcție de proporția evenimentelor similare care s-au petrecut în trecut (frecvența relativă a evenimentului A).

$$P_{emp.}(A) = \frac{numarul\ probelor\ in\ care\ evenimentul\ A\ a\ avut\ loc\ anterior}{numarul\ total\ de\ probe\ efectuate\ in\ cadrul\ exeprimentului} \quad (4.3)$$

Un alt rezultat fundamental al teoriei probabilităților este reprezentat de *Legea numerelor mari*, introdusă de matematicianul elvețian Jacob Bernoulli (1654–1705), care explică modul în care frecvențele relative ale evenimentelor se stabilizează pe măsură ce numărul de observații crește. Astfel, dacă un experiment se repetă, în aceleași condiții, de un număr suficient de mare de ori, atunci frecvența relativă a unui eveniment A prezintă o anumită stabilitate, variind în jurul probabilității de realizare a acestuia, bazată pe definiția clasică.

De exemplu, considerăm aruncarea unei monede (experimentul din figura 4.1), care poate avea două rezultate posibile (evenimentul 1 – ban și evenimentul 2 – stemă).

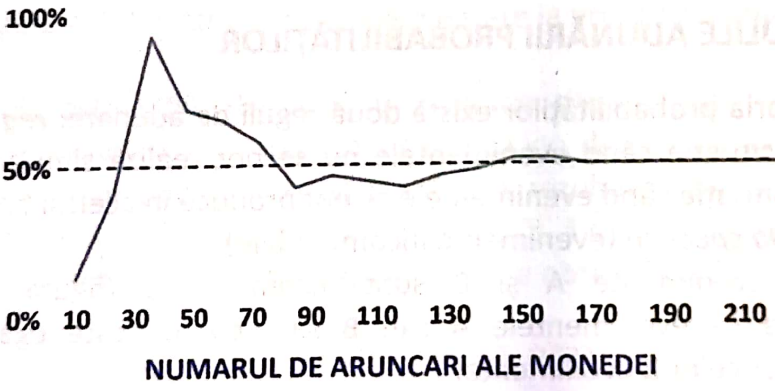


Figura 4.1. Aplicarea Legii numerelor mari în experimentul aruncării unei monede

Conform definiției clasice, probabilitatea de a obține în urma unei aruncări varianta „ban" este de 0,50 sau 50% (numărul rezultatelor favorabile evenimentului / numărul rezultatelor posibile ale experimentului).

Probabilitatea empirică de a obține varianta „cap" la o aruncare este 0 sau 1 (cele două evenimente sunt egal probabile, fiecare având aceeași șansă de a se produce). Observăm din analiza figurii 4.1, faptul că pe măsură ce efectuăm mai multe probe în cazul experimentului aruncării monedei, frecvența relativă

Scanned with OKEN Scanner

(probabilitatea empirică) a variantei „ban" se apropie de valoarea reală de 0,5 sau 50% a producerii evenimentului aferentă probabilității clasice.

Un alt exemplu interesant (Basic Statistics for Business & Economics – Lind, 2019, p. 123): Pe data de 1 februarie 2003 naveta spațială Columbia a explodat. Acesta a fost al doilea dezastru în cele 113 misiuni spațiale ale NASA. În baza acestor informații, care este probabilitatea ca o misiune viitoare să fie un succes?

$$P(A) = \frac{Numarul\ zborurilor\ finalizate\ cu\ succes}{Numarul\ total\ de\ zboruri} = \frac{113-2}{113} = \frac{111}{113} = 0{,}98$$

Vom utiliza această valoare ca o estimare a probabilității dorite. Astfel, în baza experiențelor anterioare, putem afirma că probabilitatea ca o misiune spațială viitoare să se finalizeze în siguranță este de 0,98 sau 98%.

3. *Probabilitatea subiectivă* reprezintă șansa (posibilitatea) ca un eveniment particular să aibă loc, estimată (atribuită) de către o anumită persoană în baza informațiilor disponibile în acel moment.

De exemplu, estimarea șanselor ca România să participe la Cupa mondială de fotbal, estimarea posibilității ca o persoană să fie implicată într-un accident auto în perioada următoarelor 12 luni, estimarea cursului valutar EUR/RON peste 3 luni, etc.

4.2. REGULILE ADUNĂRII PROBABILITĂȚILOR

În teoria probabilităților există două reguli de adunare: *regula specială*, aplicabilă în situația când evenimentele nu se pot realiza simultan, și *regula generală*, în situația când evenimentele se pot produce în același timp.

1. *Regula specială* (evenimente incompatibile):

Dacă 2 evenimente A și B sunt incompatibile (figura 4.2), atunci probabilitatea ca evenimentele A sau B să aibă loc este egală cu suma probabilităților celor 2 evenimente:

$$P(A\ sau\ B) = P(A \cup B) = P(A) + P(B) \qquad (4.4)$$

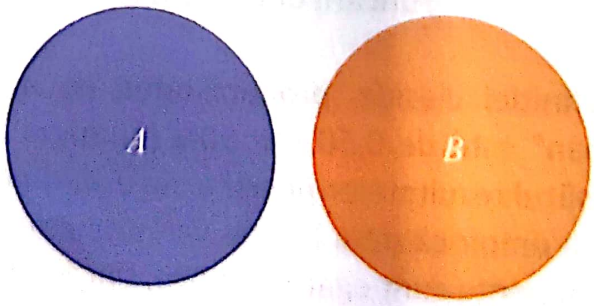


Figura 4.2. Evenimente incompatibile

Scanned with OKEN Scanner

De exemplu, la aruncarea unui zar probabilitatea de a obține o față cu 5 puncte sau una cu 6 puncte este:

$$P(5 \cup 6) = P(5) + P(6) = 1/6 + 1/6 = 2/6 = 1/3$$

Generalizând, în cazul a n evenimente incompatibile avem:

$$P(A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + P(A_3) + \cdots + P(A_n) \quad (4.5)$$

2. *Regula complementară* (evenimente contrare)

Un anumit eveniment este contrar (opus, complementar) evenimentului A, dacă realizarea sa constă în nerealizarea evenimentului A. Evenimentul contrar lui A se notează cu $\bar{A}$ (figura 4.3).

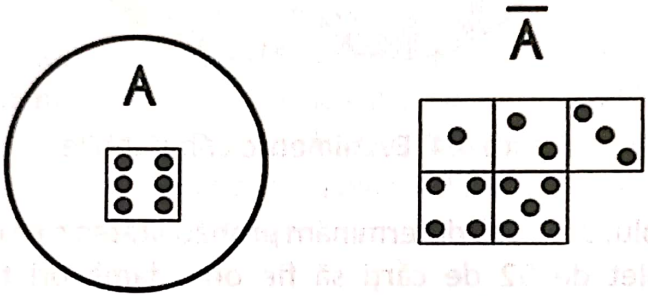


Figura 4.3. Evenimente complementare la aruncarea unui zar

Regula este utilă, deoarece în anumite situații este mai ușor să se calculeze probabilitatea ca un eveniment să aibă loc, determinând probabilitatea ca acesta să nu se întâmple și apoi scăzând rezultatul din 1. De exemplu, probabilitatea de a obține una dintre fețele 1, 2, 3, 4 sau 5 la aruncarea unui zar se poate calcula scăzând din 1 probabilitatea de a obține fața 6:

$$P(1 \cup 2 \cup 3 \cup 4 \cup 5) = P(1) + P(2) + P(3) + P(4) + P(5) =$$
$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{5}{6} = 0{,}83$$
$$P(1 \cup 2 \cup 3 \cup 4 \cup 5) = 1 - P(6) = 1 - \frac{1}{6} = \frac{5}{6} = 0{,}83$$

Indiferent de modul de calcul (ca sumă a probabilităților individuale sau ca diferență dintre 1 și probabilitatea evenimentului contrar), rezultatele obținute sunt identice.

3. *Regula generală* (evenimente compatibile)

Măsoară șansa ca două sau mai multe evenimente compatibile (realizarea unuia nu exclude realizarea celuilalt sau celorlalte) să aibă loc simultan.

Scanned with OKEN Scanner

$$P(A \text{ sau } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (4.6)$$

În relația 4.6 prezentată anterior, $P(A \cap B) = P(A \text{ și } B)$ reprezintă probabilitatea realizării simultane a celor două evenimente A și B sau șansa ca ele să se întâmple concomitent (intersecția evenimentelor A și B).

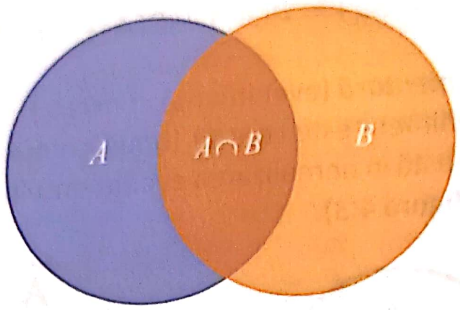


Figura 4.4. Evenimente compatibile

De exemplu, dorim să determinăm probabilitatea ca o carte extrasă dintr-un pachet complet de 52 de cărți să fie ori damă ori treflă (evenimente compatibile). Astfel, avem:

$$P\,(Dama) = 4 / 52$$
$$P\,(Trefla) = 13 / 52$$
$$P\,(Damă\ de\ Trefla) = 1 / 52$$

Aplicăm regula generală a adunării probabilităților:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52} = \frac{16}{52} = 0{,}31$$

Probabilitatea realizării simultane a celor două evenimente trebuie să fie scăzută din suma probabilităților individuale, deoarece, așa cum se observă în figura 4.3, aceasta este inclusă în fiecare eveniment și ar exista riscul să fie cuprinsă de două ori în rezultatul $P(A \cup B)$, conducând astfel la denaturarea lui.

4.3. REGULILE ÎNMULȚIRII PROBABILITĂȚILOR

Acest set de reguli se utilizează în situația determinării probabilității ca două evenimente să aibă loc simultan. De exemplu, care este probabilitatea ca o persoană care a contactat un virus (evenimentul A) să dezvolte o formă ușoară a

Scanned with OKEN Scanner

bolii (evenimentul B)? Înmulțirea probabilităților este tratată diferit în situația când evenimentele sunt independente (regula specială) sau dependente (regula generală).

1. *Regula specială* (evenimente independente)

Este folosită în situația când evenimentele studiate sunt independente (realizarea unuia nu influențează probabilitatea realizării celuilalt).

$$P(A\ si\ B) = P(A \cap B) = P(A) \cdot P(B) \qquad (4.7)$$

În cazul a 3 evenimente independente A, B și C avem:

$$P(A\ si\ B\ si\ C) = P(A \cap B \cap C) = P(A) \cdot P(B) \cdot P(C) \qquad (4.8)$$

Exemplu: două fabrici de medicamente produc același tip de medicament. Probabilitățile ca medicamentul să corespundă standardului sunt 0.98 la prima fabrică, respectiv 0.96 la cea de a doua (evenimente independente). Se ia pentru încercare, cate o cutie de la fiecare fabrică. Care este probabilitatea ca ambele cutii să fie corespunzătoare?

$$P(A \cap B) = P(A) \cdot P(B) = 0{,}98 \cdot 0{,}96 = 0{,}94$$

2. *Regula generală* (evenimente dependente)

Este utilizată atunci când evenimentele studiate sunt dependente (probabilitatea unuia este influențată de evenimentele anterioare – dacă acestea s-au produs sau nu).

Exemplu: presupunem că avem 10 cutii de suc în frigider, 7 de portocale și 3 de ananas. O cutie este extrasă aleatoriu din frigider. Probabilitatea unui suc de portocale este 7 / 10 și a unui suc de ananas 3 / 10. Ulterior, o a doua cutie este scoasă din frigider, fără a o așeza înapoi pe prima extrasă. Probabilitatea celei de a doua cutii de suc depinde de tipul primei cutii selectate (de portocale sau de ananas).

Astfel, probabilitatea ca a doua cutie de suc să fie de ananas este:

- 2 / 9 dacă prima cutie de suc a fost tot de ananas;
- 3 / 9 dacă prima cutie de suc a fost de portocale.

Această probabilitate (2 / 9 sau 3 / 9) reprezintă o probabilitate condiționată, întrucât valoarea ei depinde de rezultatul evenimentului anterior (ce tip de cutie de suc a fost inițial extrasă din frigider).

Probabilitatea condiționată reprezintă astfel probabilitatea realizării unui eveniment prezent (B) ținând cont de influența exercitată de un alt eveniment petrecut anterior (A) și se notează cu $P_A(B)$ sau $P(B|A)$

Scanned with OKEN Scanner

Prin urmare, în cazul a două evenimente dependente A și B regula înmulțirii probabilităților devine:

$$P(A\ si\ B) = P(A \cap B) = P(A) \cdot P(B|A) \tag{4.9}$$

4.4. PRINCIPIILE NUMĂRĂRII

În situația în care numărul rezultatelor posibile în cadrul unui experiment este redus, este relativ ușor să le numărăm direct. De exemplu, în cazul aruncării unui zar avem șase rezultate posibile: 1, 2, 3, 4, 5 și 6. În general însă, ne confruntăm cu un număr ridicat de rezultate posibile pentru un anumit experiment. Astfel, în vederea alegerii ținutei de prezentare la un interviu o persoană are de ales între 3 sacouri, 4 cămăși, 5 cravate și 2 perechi de pantofi. În câte modalități posibile se poate îmbrăca persoana pentru interviu?

Principiul fundamental al numărării: dacă un eveniment are m rezultate posibile și alt eveniment are n rezultate posibile, atunci există $m \cdot n$ rezultate posibile pentru cele 2 evenimente considerate împreună.

Relația se poate extinde pentru mai mult de 2 evenimente, având fiecare $m_1, m_2, m_3, \ldots, m_k$ rezultate posibile: numărul total de rezultate posibile ale celor k evenimente considerate împreună va fi $m_1 \cdot m_2 \cdot m_3 \cdot \ldots \cdot m_k$ rezultate.

Revenind la exemplul anterior, numărul total de modalități posibile în care se poate îmbrăca respectiva persoană este $3 \cdot 4 \cdot 5 \cdot 2 = 120$ ținute complete.

Aranjamentele (permutările) reprezintă numărul total de submulțimi formate din k elemente ordonate ale unei mulțimi A compusă din n elemente $(0 \leq k \leq n, n \epsilon N)$. Se determină prin aplicarea formulei:

$$A_n^k = \frac{n!}{(n-k)!} \tag{4.10}$$

Să presupunem, de exemplu că dorim să alegem toate grupurile de 2 litere dintr-o mulțime care conține 4 litere: A, B, C și D (figura 4.5).

$$A_4^2 = \frac{4!}{(4-2)!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1} = 12\ grupuri$$

Deoarece ordinea contează, grupurile AB și BA, AC și CA, AD și DA, BC și CB, BD și DB, CD și DC sunt diferite.

Scanned with OKEN Scanner

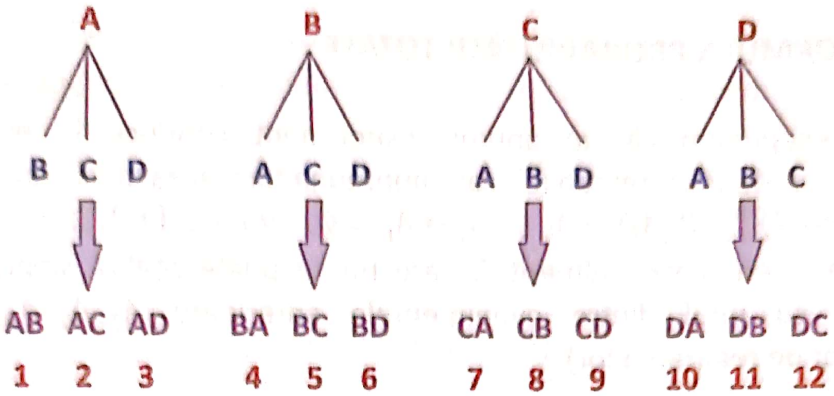


Figura 4.5. Grupurile de 2 litere dintr-o mulțime care conține 4 litere

Dacă ordinea obiectelor selectate nu este importantă, selecția aranjamentelor (grupurilor) îmbracă forma *combinărilor*. În mod logic, numărul combinărilor este întotdeauna mai mic decât cel al aranjamentelor (permutărilor).

$$C_n^k = \frac{n!}{k! \cdot (n-k)!} \tag{4.11}$$

Revenind la exemplul anterior prezentat în cazul permutărilor, deoarece ordinea literelor nu este importantă, grupurile AB și BA, AC și CA, AD și DA, BC și CB, BD și DB, CD și DC sunt identice și vor fi luate în considerare doar o singură dată. Astfel, numărul de variante s-a înjumătățit:

$$C_4^2 = \frac{4!}{2! \cdot (4-2)!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 2 \cdot 1} = 6\ grupuri$$

Un alt exemplu: pentru a evalua nivelul de satisfacție al celor 100 de pacienți din cadrul unei secții a unui spital universitar cu privire la condițiile oferite pe perioada internării, vom extrage aleatoriu un eșantion de 8 pacienți.

$$C_{100}^8 = \frac{100!}{8! \cdot (100-8)!} = 186.087.894.300\ esantioane$$

Acest lot de pacienți va fi unul dintre cele 186.087.894.300 de eșantioane, care pot fi obținute dacă am extrage la un moment dat toate grupurile posibile de 8 persoane din populația de referință de 100 de pacienți (C_{100}^8).

Scanned with OKEN Scanner

4.5. FORMULA PROBABILITĂȚII TOTALE

Presupunem că un anumit experiment conduce la evenimentele $A_1, A_2, A_3, \dots, A_n$ (incompatibile), care împreună formează un sistem complet de evenimente ($\sum_{i=1}^{n} P(A_i) = 1$ și $A_i \cap A_j = \emptyset$, $\forall\, i,j \in \{1,2,\dots,n\},\ i \neq j$). De asemenea, avem un eveniment X care nu se poate realiza singur, ci numai împreună cu unul dintre evenimentele anterioare $A_1, A_2, A_3, \dots, A_n$ (sau condiționat de realizarea lor).

Prin urmare avem:

$$X = (X \cap A_1) \cup (X \cap A_2) \cup (X \cap A_3) \cup \dots \cup (X \cap A_n)$$

Deoarece aceste evenimente sunt incompatibile două câte două, conform regulii speciale de adunare a probabilităților rezultă:

$$P(X) = P(X \cap A_1) + P(X \cap A_2) + P(X \cap A_3) + \cdots + P(X \cap A_n) \qquad (4.12)$$

Pentru fiecare termen $P(X \cap A_i)$ din dreapta șirului folosim formula (4.9) a probabilităților condiționate:

$$\begin{aligned} P(X \cap A_1) &= P(A_1) \cdot P(X|A_1) \\ P(X \cap A_2) &= P(A_2) \cdot P(X|A_2) \\ &\dots\dots\dots\dots\dots\dots\dots\dots \\ P(X \cap A_n) &= P(A_n) \cdot P(X|A_n) \end{aligned} \qquad (4.13)$$

Înlocuind în formula 4.12 termenii dezvoltați sub forma probabilităților condiționate (4.13), obținem formula probabilității totale:

$$P(X) = P(A_1) \cdot P(X|A_1) + P(A_2) \cdot P(X|A_2) + \cdots + P(A_n) \cdot P(X|A_n) \qquad (4.14)$$

Exemplul 1: într-o farmacie se găsesc medicamente de același fel, provenite de la trei depozite de medicamente (A, B și C). Se știe că 25% din totalul medicamentelor provin din primul depozit, 35% din al doilea depozit, iar 40% din al treilea depozit. Procentele de rebuturi aferente fiecărui depozit de medicamente sunt 2%, 3% și respectiv 1% pentru fiecare dintre ele. Să se calculeze probabilitatea ca luând un medicament la întâmplare din farmacie, acesta să fie necorespunzător calitativ.

A, B și C reprezintă evenimentele ca medicamentul ales să provină de la unul dintre cele 3 depozite și X ca medicamentul să fie necorespunzător calitativ.

Scanned with OKEN Scanner

Deoarece medicamentul necorespunzător calitativ poate proveni numai de la unul dintre cele 3 depozite, rezultă faptul că evenimentul X aferent nu se poate realiza singur, ci numai împreună cu A, B sau C. Astfel, avem:

$$P(A) = 0{,}25 \qquad P(X \mid A) = 0{,}02$$
$$P(B) = 0{,}35 \qquad P(X \mid B) = 0{,}03$$
$$P(C) = 0{,}40 \qquad P(X \mid C) = 0{,}01$$

Utilizăm relația probabilității totale (formula 4.14):

$$P(X) = P(A) \cdot P(X|A) + P(B) \cdot P(X|B) + P(C) \cdot P(X|C) =$$
$$= 0{,}25 \cdot 0{,}02 + 0{,}35 \cdot 0{,}03 + 0{,}40 \cdot 0{,}01 = 0{,}0195$$

Putem concluziona faptul că, la fiecare 10000 de medicamente achiziționate de la cele 3 depozite, în medie 195 sunt necorespunzătoare calitativ.
Exemplul 2: dintr-un grup de copii, 60% sunt vaccinați împotriva unei afecțiuni. Daca un copil este vaccinat, probabilitatea de a contacta afecțiunea este de 0,001. Pentru cei nevaccinați probabilitatea este de 10 ori mai mare, respectiv 0,01. Dacă se alege un copil la întâmplare, care este probabilitatea ca acesta să contacteze afecțiunea?

Se presupun cunoscute următoarele informații:
60% dintre copii sunt vaccinați (V): $P(V) = 0{,}6$ $P(A \mid V) = 0{,}001$
40% dintre copii sunt nevaccinați (N): $P(N) = 0{,}4$ $P(A \mid N) = 0{,}01$

Utilizăm relația probabilității totale (formula 4.14):

$$P(A) = P(V) \cdot P(A \mid V) + P(N) \cdot P(A|N) =$$
$$= 0{,}6 \cdot 0{,}001 + 0{,}4 \cdot 0{,}01 = 0{,}0046$$

Prin urmare, probabilitatea ca un copil ales la întâmplare din acel grup (vaccinat sau nevaccinat) să contacteze afecțiunea este 0,46%.

4.6. TEOREMA LUI BAYES

Teorema lui Bayes este utilizată pentru a calcula probabilitatea unui eveniment, având informații anterioare despre acel eveniment.

Considerăm $D_1, D_2, D_3, \ldots, D_n$ un sistem complet de evenimente ($\sum_{i=1}^{n} P(D_i) = 1$), iar S un alt eveniment pentru care toate probabilitățile condiționate $P(S|D_1), P(S|D_2), P(S|D_3), \ldots, P(S|D_n)$ sunt cunoscute, atunci probabilitățile condiționate inverse $P(D_i|S)$ se obțin cu formula:

Scanned with OKEN Scanner

$$P(D_i|S) = \frac{P(D_i) \cdot P(S|D_i)}{\sum_{i=1}^{n} P(D_i) \cdot P(S|D_i)} \quad (4.15)$$

Probabilitățile $P(D_i)$ sunt cunoscute din cercetări anterioare sau pot fi estimate prin diverse metode (utilizarea datelor de la recensământ). La numărătorul formulei lui Bayes avem probabilitatea condițională, iar la numitor probabilitatea totală.

Revenim la datele prezentate anterior în exemplul 2 din subcapitolul 4.5. Întrebarea care se pune acum este: dacă un copil ales la întâmplare a contractat afecțiunea, care este probabilitatea ca el să fi fost anterior vaccinat? Dar nevaccinat?

Vom aplica Teorema lui Bayes. Avem informații prealabile, adică știm că copilul a contractat afecțiunea. Astfel, dorim să știm care este probabilitatea ca acesta să fi fost vaccinat sau nevaccinat cunoscând faptul că a contractat afecțiunea.

Se presupun cunoscute următoarele informații:

60% dintre copii sunt vaccinați (V): $P(V) = 0{,}60$ $\quad P(A \mid V) = 0{,}001$

40% dintre copii sunt nevaccinați (N): $P(N) = 0{,}40$ $\quad P(A \mid N) = 0{,}01$

Probabilitatea ca un copil ales la întâmplare să contracteze afecțiunea, indiferent dacă a fost sau nu anterior vaccinat, a fost determinată anterior, în subcapitolul 4.5, prin utilizarea formulei probabilității totale:

$$P(A) = P(V) \cdot P(A|V) + P(N) \cdot P(A|N) = 0{,}0046$$

În continuare, vom utiliza formula lui Bayes (4.15), pentru ambele situații:

- Probabilitatea ca un copil *să fie vaccinat* știind că a contractat afecțiunea este:

$$P(V|A) = P(V \cap A) / P(A) = (P(V) * P(A/V)) / P(A) = \\ = 0{,}60 * 0{,}001 / 0{,}0046 = 0{,}13$$

- Probabilitatea ca un copil *să fie nevaccinat* știind că a contractat afecțiunea este:

$$P(N|A) = P(N \cap A) / P(A) = (P(N) * P(A/N)) / P(A) = \\ = 0{,}40 * 0{,}01 / 0{,}0046 = 0{,}87$$

În concluzie, probabilitatea ca un copil ales la întâmplare din grup care a contractat afecțiunea să fie vaccinat este de doar 0,13 sau 13%, aproximativ de 6,7 ori mai mică, comparativ cu situația în care copilul nu a fost vaccinat (0,87 sau 87%).

Scanned with OKEN Scanner

Capitolul. 5

DISTRIBUȚII PROBABILISTICE DISCRETE ȘI CONTINUE

SUBIECTE:

5.1. Definiția unei distribuții probabilistice
5.2. Variabilele aleatoare discrete și continue
5.3. Media aritmetică și varianța unei distribuții probabilistice discrete
5.4. Media aritmetică și varianța unei distribuții probabilistice continue
5.5. Distribuții probabilistice discrete - binomială, hipergeometrică, Poisson
5.6. Distribuții probabilistice continue – uniformă și normală
5.7. Distribuția normală standard

La finalizarea acestui capitol, studentul va fi capabil să:
O5-1: să identifice caracteristicile unei distribuții de probabilitate
O5-2: să facă distincția între o variabilă aleatoare discretă și una continuă
O5-3: să calculeze media, varianța și deviația standard în cazul unei distribuții discrete
O5-4: să explice caracteristicile unei anumite distribuții discrete (binomială, hipergeometrică sau Poisson) și să le aplice în vederea calculării probabilităților
O5-5: să descrie proprietățile distribuțiilor probabilistice continue (uniformă și normală) și să le utilizeze pentru determinarea probabilităților
O5-6: să descrie proprietățile distribuției normale standard (redusă) și să le utilizeze în calculul probabilităților

5.1. DEFINIȚIA UNEI DISTRIBUȚII PROBABILISTICE

În cadrul capitolului anterior au fost introduse conceptele de bază ale teoriei probabilităților alături de metodele principale de determinare a probabilității unui anumit eveniment. În continuare, vom utiliza aceste noțiuni, explorând noi modalități de calcul ale probabilității unui eveniment în condiții mai complexe. Astfel, vom vedea imediat faptul că relația dintre valorile pe care le poate lua o variabilă statistică aleatoare și probabilitățile apariției acestora pot fi sintetizate sub forma unei *„distribuții probabilistice"*.

Cunoașterea distribuției probabilistice a unei variabile aleatoare reprezintă pentru un cercetător un instrument important în centralizarea și descrierea unui set de date statistice în vederea formulării unor concluzii

Scanned with OKEN Scanner

pertinente referitoare la o anumită populație (colectivitate generală) pe baza datelor obținute la nivelul unui eșantion reprezentativ extras din aceasta.

O distribuție probabilistică este asemănătoare cu o distribuție a frecvențelor relative (figura 5.1), cu precizarea însă că, în loc să descrie evenimente care s-au derulat anterior, este utilizată în vederea obținerii unor estimări a șansei de producere în viitor a unor evenimente.

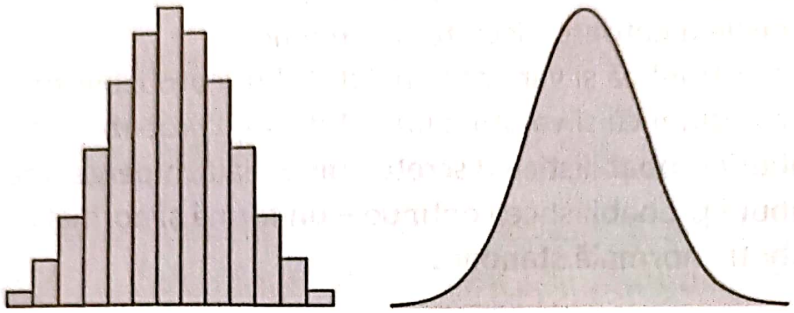


Figura 5.1. Distribuții probabilistice

Distribuția probabilistică reprezintă o listă (tabel, grafic) care cuprinde întreaga plajă de rezultate posibile ale unui experiment împreună cu toate nivelurile probabilităților asociate fiecărui rezultat în parte.

Proprietățile unei distribuții probabilistice

- Probabilitatea realizării fiecărui eveniment este cuprinsă în intervalul [0,1];
- Evenimentele sunt incompatibile (se exclud reciproc);
- Evenimentele formează un sistem complet de evenimente, astfel încât suma probabilităților lor de realizare este întotdeauna 1.

Tabelul 5.1. Rezultatele experimentului „aruncarea de 3 ori a unei monede"

Rezultate posibile	Aruncarea monedei			Numărul de variante „cap"
	prima	a doua	a treia	
1	P	P	P	0
2	P	P	C	1
3	P	C	P	1
4	P	C	C	2
5	C	P	P	1
6	C	P	C	2
7	C	C	P	2
8	C	C	C	3

Scanned with OKEN Scanner

Pentru a exemplifica, presupunem că suntem interesați în realizarea unui experiment care presupune aflarea numărului de variante „cap" care pot apare în situația aruncării de 3 ori a unei monede (tabelul 5.1). Aplicăm regula multiplicării și avem $2 \cdot 2 \cdot 2 = 8$ rezultate posibile în derularea experimentului, fiecare cu un număr diferit de apariții. Centralizând, obținem doar 4 rezultate diferite (evenimente): să avem 0, 1 ,2 sau 3 variante „cap" în urma aruncării monedei (coloana 5 din tabelul 5.1).

În continuare realizăm sub formă tabelară distribuția probabilistică pentru cele 4 evenimente (0, 1, 2, 3 variante „cap") rezultate în urma aruncării de trei ori a unei monede (tabelul 5.2). Deoarece unul dintre aceste rezultate trebuie să aibă loc, suma probabilităților tuturor evenimentelor posibile este egală cu 1.

Tabelul 5.2. Distribuția probabilistică a experimentului „aruncarea unei monede"

Numărul de variante „cap" x_i	**Probabilitatea de apariție** $P(x_i)$
0	1 / 8 = 0,125
1	3 / 8 = 0,375
2	3 / 8 = 0,375
3	1 / 8 = 0,125
Total	**8 / 8 = 1,000**

Informațiile privind cele 4 evenimente (x_i) prezentate în tabelul 5.2 împreună cu probabilitățile aferente acestora $P(x_i)$ pot fi înfățișate și sub formă grafică:

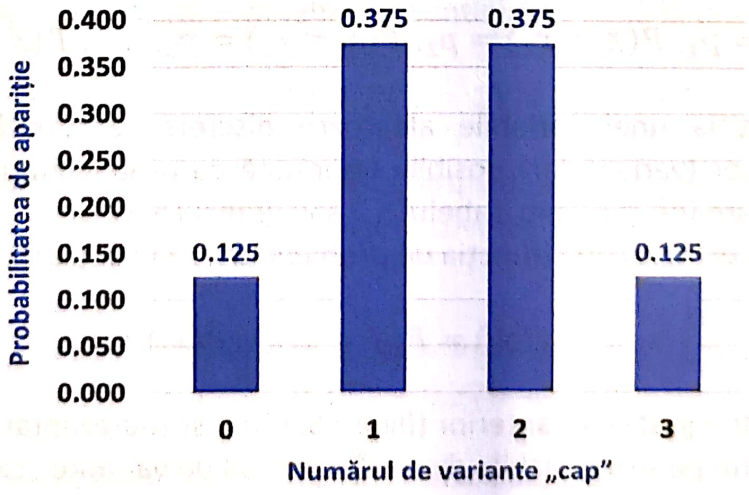


Figura 5.2. Distribuția probabilistică a experimentului „aruncarea unei monede"

Scanned with OKEN Scanner

Analizând distribuția probabilistică reprezentată în figura 5.2, observăm faptul că probabilitățile cele mai mari de apariție sunt înregistrate de categoriile 1 și 2 variante „cap", fiecare având 37,5%. Probabilitățile extreme, respectiv nicio variantă „cap" sau 3 variante „cap", înregistrează valori mult mai reduse, fiecare fiind de 12,5%.

5.2. VARIABILELE ALEATOARE DISCRETE ȘI CONTINUE

În orice experiment care implică șansa, rezultatele apar în mod întâmplător. De exemplu, aruncarea unui zar este un eveniment, în care oricare dintre cele șase fețe poate să apară. O parte dintre experimente pot avea rezultatele evidențiate numeric fiind exprimate sub forma unor variabile cantitative (greutatea, înălțimea, glicemia, colesterolul, numărul de copii, etc.) iar altă parte pot genera rezultate sub forma unor tipuri calitative (categorii) diferite fiind exprimate sub forma unor variabile calitative (grupa de sânge, genul, starea civilă, culoarea ochilor, mediul social, etc.).

O *variabilă aleatoare* reprezintă o variabilă măsurată sau observată ca rezultat al unui experiment, care în mod cu totul întâmplător poate lua diferite valori. Există două tipuri de variabile aleatoare: *discrete* și *continue*. Variabilele discrete pot lua doar un număr limitat de valori izolate (numărul de copii dintr-o familie, notele studenților la un examen). Variabilele continue pot lua însă absolut orice valoare în cadrul unui anumit interval (temperatura corpului uman, înălțimea, greutatea, glicemia, colesterolul, etc.).

1. Variabilele aleatoare discrete

Considerăm cazul unei variabile aleatoare X. În urma realizării unui experiment se obțin valorile x_1, x_2, x_3, ..., x_n cu probabilitățile de apariție:

$$P(X = x_1) = p_1,\ P(X = x_2) = p_2,\ P(X = x_3) = p_3,\ \ldots,\ P(X = x_n) = p_n.$$

Repartiția unei variabile aleatoare discrete reprezintă enumerarea tuturor valorilor (variantelor) posibile împreună cu probabilitățile de obținere corespunzătoare (de exemplu, tabelul 5.2 sau graficul 5.2).

Legea de repartiție (funcția de probabilitate) a unei variabile aleatoare X este:

$$f(x_i) = P(X = x_i) = P(x_i) = p_i \qquad (5.1.)$$

În cadrul graficului anterior (figura 5.2) a fost reprezentată grafic funcția de probabilitate pentru variabila discretă numărul de variante „cap" rezultate în urma aruncării de trei ori a unei monede.

Scanned with OKEN Scanner

În unele situații este necesară determinarea probabilității ca valoarea variabilei aleatoare să fie mai mică decât un anumit nivel prag x_i.

Vom determina astfel, *funcția de repartiție*:

$$F(x_i) = P(X \leq x_i) = \sum_{i=1}^{k} P(X = x_i) = \sum_{i=1}^{k} p_i \tag{5.2}$$

Continuând pe linia aceluiași exemplu prezentat în tabelul 5.2 a fost construită funcția de repartiție a variabilei discrete numărul de variante „cap" rezultate în urma aruncării de trei ori a unei monede (figura 5.3):

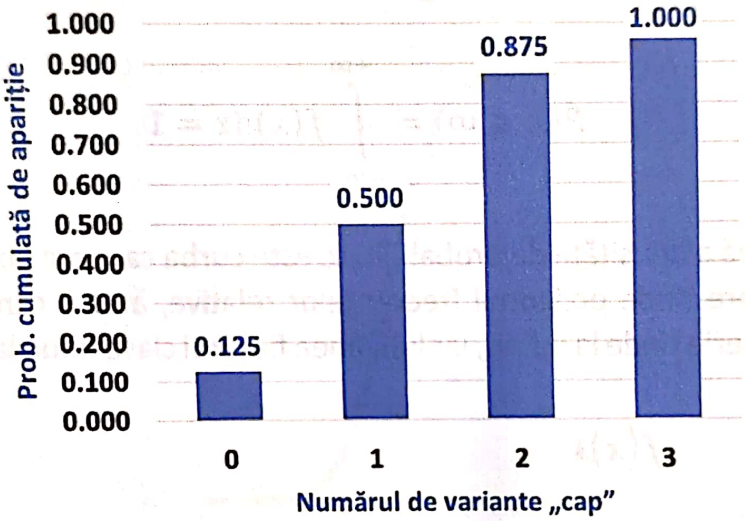


Figura 5.3. Funcția de repartiție

În figura 5.3 probabilitățile au fost cumulate treptat, începând cu cea specifică primei variante (0) a variabilei și continuând până la ultima (3). Astfel:

$$F(1) = P(X \leq 1) = P(X = 0) + P(X = 1) = 0{,}125 + 0{,}375 = 0{,}500$$

reprezintă probabilitatea de 50%, ca la aruncarea de trei ori a monedei să obținem 0 sau 1 variante „cap", sau cu alte cuvinte cel puțin o variantă „cap".

În mod similar:

$$F(2) = P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = \\ = 0{,}125 + 0{,}375 + 0{,}375 = 0{,}875$$

reprezintă probabilitatea de 87,5% ca la aruncarea de trei ori a monedei să obținem 0, 1 sau 2 variante „cap", sau cu alte cuvinte cel puțin două variante „cap".

Scanned with OKEN Scanner

2. *Variabilele aleatoare continue*

Acestea pot lua un număr „infinit" de valori în cadrul unui anumit interval. Din acest considerent, evidențierea sub formă tabelară a tuturor valorilor pe care le poate avea o variabilă aleatoare continuă nu este posibilă. Prin urmare, se impune utilizarea probabilității evenimentului $X < a$, operațiile fiind efectuate de fiecare dată asupra unor intervale de valori și nu asupra unor valori fixe.

În locul funcției de probabilitate, specifică variabilelor discrete, se va utiliza *funcția densitate de probabilitate*.

Suprafața totală cuprinsă între curbă și axa orizontală este egală cu 1, reprezentând probabilitatea ca x să se situeze în intervalul de valori $(-\infty, +\infty)$ sau probabilitatea evenimentului sigur:

$$P(X < \infty) = \int_{-\infty}^{+\infty} f(x)dx = 1 \tag{5.3}$$

O curbă a densității de probabilitate este curba care are același aspect cu curba către care tinde poligonul frecvențelor relative, atunci când numărul de valori dintr-o serie tinde la infinit, iar lungimea fiecărei clase tinde la 0 (figura 5.4).

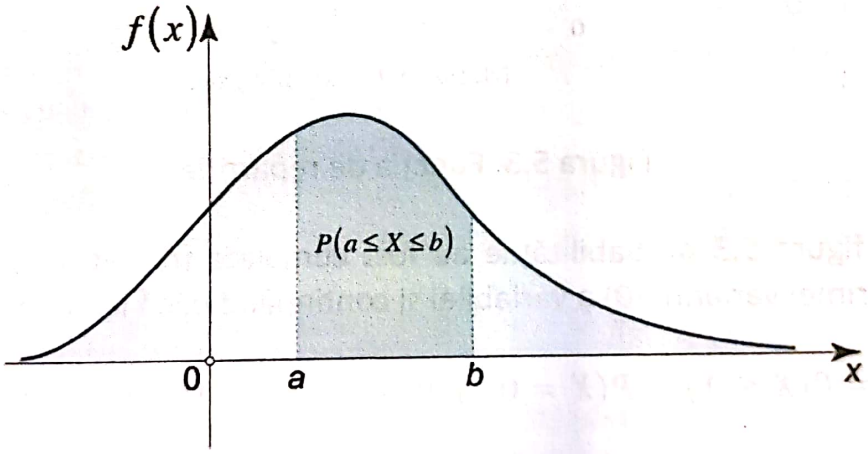


Figura 5.4. Funcția densitate de probabilitate a unei variabile continue

Probabilitatea ca variabila aleatoare continuă să ia valori în intervalul (a, b) este egală cu suprafața cuprinsă între cele 2 valori (figura 5.4). Astfel:

$$P(a,b) = P(a \le x \le b) = P\big((x \le b) \cap (x \ge a)\big) = P(x \le b) - P(x \le a)$$

$$P(a,b) = \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx = \int_{a}^{b} f(x)dx \tag{5.4}$$

Scanned with OKEN Scanner

În situația când $a = b$, probabilitatea va înregistra valoarea zero, evidențiind faptul că posibilitatea ca variabila continuă să înregistreze o valoare fixă este atât de mică în cazul unui interval care conține un număr infinit de valori, încât poate fi asimilată valorii nule.

Funcția de repartiție sau funcția de distribuție cumulativă a unei variabile aleatoare continue este o funcție matematică ce descrie probabilitatea ca variabila aleatoare X să ia valori mai mici sau egale decât un anumit număr real x. Aceasta este definită astfel:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx \tag{5.5}$$

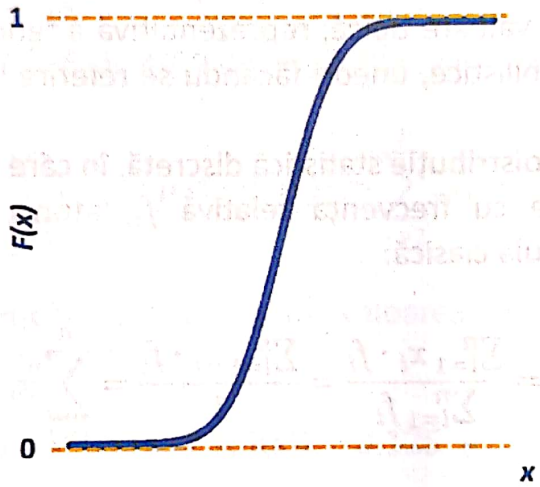


Figura 5.5. Funcția de repartiție a unei variabile continue

Funcția de repartiție a unei variabile aleatoare continue X este monoton crescătoare (figura 5.5): pentru oricare valori $x_1 \leq x_2$ avem $F(x_1) \leq F(x_2)$. Are asimptote bine definite, respectiv dreptele: $F(x) = 0$ și $F(x) = 1$.

Funcția de repartiție are multiple valențe practice, fiind utilă pentru:

- calcularea probabilității ca o variabilă aleatoare X să ia valori mai mici sau egale decât un anumit prag x;
- caracterizarea completă a unei distribuții discrete sau continue;
- compararea a două distribuții diferite;
- determinarea valorii tendinței centrale: mediana distribuției este valoarea x pentru care $F(x) = 0{,}50$;
- determinarea funcției densității de probabilitate prin derivarea funcției de repartiție;
- modelarea fenomenelor reale în medicină, de exemplu analiza timpului de supraviețuire în cazul pacienților diagnosticați cu un

Scanned with OKEN Scanner

anumit tip de cancer – funcția de repartiție a variabilei X, timpul în luni până la deces, ne arată în această situație probabilitatea ca un pacient să decedeze înainte de momentul de timp x. Dacă $F(10) = 0{,}25$ înseamnă că un sfert dintre pacienți decedează în primele 10 luni. Pot fi construite în acest context, funcții diferite de repartiție, pentru pacienții care urmează două tipuri de tratament: chimioterapie standard vs. terapie țintită.

5.3. MEDIA ARITMETICĂ ȘI VARIANȚA UNEI DISTRIBUȚII PROBABILISTICE DISCRETE

1. Media aritmetică a unei distribuții discrete

Media este o valoare tipică, reprezentativă a tendinței centrale în cazul unei distribuții probabilistice, uneori făcându-se referire la ea ca fiind „valoarea așteptată".

Dacă avem o distribuție statistică discretă, în care fiecare nivel individual al variabilei x_i apare cu frecvența relativă f_i, atunci media aritmetică se calculează după formula clasică:

$$\mu = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{1} = \sum_{i=1}^{n} x_i \cdot f_i \tag{5.6}$$

Conform legii numerelor mari, dacă un experiment se repetă, în aceleași condiții, de un număr suficient de mare de ori, atunci *frecvența relativă a unui eveniment* f_i prezintă o anumită *stabilitate*, variind în jurul probabilității de realizare a acestuia p_i.

Rezultă imediat faptul că, *media aritmetică a unei distribuții probabilistice discrete* se poate determina ca o medie aritmetică ponderată, în care toate valorile posibile ale unei variabile aleatoare x_i sunt multiplicate cu probabilitățile corespunzătoare de apariție ale lor $P(x_i) = p_i$:

$$\mu = \sum_{i=1}^{n} [x_i \cdot P(x_i)] = \sum_{i=1}^{n} x_i \cdot p_i \tag{5.7}$$

Vom determina valoarea mediei numărului de variante „cap" conform datelor din tabelul 5.2, în care a fost prezentată inițial distribuția probabilistică a rezultatelor experimentului „aruncarea unei monede":

Scanned with OKEN Scanner

$$\mu = \sum_{i=1}^{4} x_i \cdot p_i = 0 \cdot 0{,}125 + 1 \cdot 0{,}375 + 2 \cdot 0{,}375 + 3 \cdot 0{,}125 = 1{,}500$$

Astfel, la aruncarea de trei ori a unei monede, varianta „cap" va apare în medie de „1,5" ori sau, cu alte cuvinte, vom putea spera să obținem de „1,5" ori varianta „cap" în cadrul acestui experiment.

2. *Varianța unei distribuții discrete*

Media este o valoare reprezentativă spre care tind toate valorile individuale ale unei distribuții probabilistice, cunoscută din acest motiv sub numele de „speranța matematică" a unei variabile. Cu toate acestea, media nu poate cuantifica nivelul împrăștierii (variației) dintr-o distribuție statistică, acesta fiind evidențiat de varianță.

Formula de calcul a *varianței în cazul unei distribuții probabilistice discrete* este:

$$\sigma^2 = \sum_{i=1}^{n} [(x_i - \mu)^2 \cdot P(x_i)] = \sum_{i=1}^{n} (x_i - \mu)^2 \cdot p_i \qquad (5.8)$$

Utilizând formula, 5.8 calculăm valoarea varianței pentru datele prezentate anterior în tabelul 5.2:

$$\sigma^2 = \sum_{i=1}^{4} (x_i - 1{,}5)^2 \cdot p_i = (0 - 1{,}500)^2 \cdot 0{,}125 + (1 - 1{,}500)^2 \cdot 0{,}375 + \\ + (2 - 1{,}500)^2 \cdot 0{,}375 + (3 - 1{,}500)^2 \cdot 0{,}125 = 0.750$$

Rezultă că deviația standard a variabilei va fi:

$$\sigma = \sqrt{\sigma^2} = \sqrt{0{,}750} = 0{,}866$$

În concluzie, numărul de variante „cap" se va abate în medie de la nivelul central $\mu = 1{,}500$, determinat anterior, cu $\pm$ 0,866.

5.4. MEDIA ARITMETICĂ ȘI VARIANȚA UNEI DISTRIBUȚII PROBABILISTICE CONTINUE

1. *Media aritmetică a unei distribuții continue*

Este un indicator reprezentativ și mai poartă numele de speranță matematică, valoare așteptată sau medie teoretică. În cazul variabilelor continue, media aritmetică se calculează după formula:

Scanned with OKEN Scanner

$$\mu = M[X] = \int_{-\infty}^{+\infty} x \cdot f(x)dx \tag{5.9}$$

2. *Varianța unei distribuții discrete*

Dispersia sau varianța, în calitate de indicator sintetic al împrăștierii variantelor unei variabile continue în jurul mediei aritmetice se determină astfel:

$$\sigma^2 = D[X] = M[(x-\mu)^2] = \int_{-\infty}^{+\infty} (x-\mu)^2 \cdot f(x)dx \tag{5.10}$$

5.5. DISTRIBUȚII PROBABILISTICE DISCRETE - BINOMIALĂ, HIPERGEOMETRICĂ, POISSON

1. DISTRIBUȚIA BINOMIALĂ

Distribuția binomială, introdusă de către matematicianul elvețian Jakob Bernoulli (1655-1705), fost profesor la Universitatea din Basel, este una dintre cele mai des întâlnite distribuții probabilistice discrete (reprezintă modelul extragerii bilelor din urnă cu revenire – compoziția urnei rămâne nemodificată pe tot parcursul experimentului; astfel fiecare unitate aleasă și inclusă în eșantion este returnată în populația de referință, înaintea selectării următoarei unități).

Modelul probabilistic binomial presupune îndeplinirea următoarelor condiții:

- Fiecare experiment are doar două rezultate posibile, complementare, denumite generic succes A și insucces $\bar{A}$ (de ex. rezultatul unui proces de vânzare prin telefon: un client poate cumpăra sau nu respectivul produs, dar cele 2 rezultate nu pot coexista);
- Probabilitățile evenimentului A și al opusului său $\bar{A}$ rămân constante pe parcursul derulării experimentelor (de ex. în cadrul unui test experiment cu 20 de întrebări, care au sugerate 4 răspunsuri posibile, din care doar unul este corect, avem 20 de probe, iar probabilitatea de a selecta în mod aleatoriu răspunsul corect pentru fiecare întrebare este mereu aceeași: 0,25);
- Fiecare probă din cadrul experimentului este independentă de toate celelalte, rezultatul ei neafectând în niciun fel restul probelor (răspunsul corect sau greșit la o întrebare nu influențează în niciun fel răspunsurile la toate celelalte întrebări din test).

Notăm probabilitatea unui succes $P(A) = p$ și probabilitatea unui insucces $P(\bar{A}) = q$. Astfel, probabilitatea ca efectuând de n ori experiența, să se

Scanned with OKEN Scanner

obțină de k ori un succes (evenimentul A) și implicit de $n - k$ ori un insucces (evenimentul $\bar{A}$) este:

$$P(X = x_k) = C_n^k \cdot p^k \cdot q^{n-k} \tag{5.11}$$

Evenimentele A și $\bar{A}$ sunt independente și conform regulii speciale de înmulțire a probabilităților vom avea:

$$\begin{gathered} P = P(A \cap A \cap A \cap \ldots \cap A \cap \bar{A} \cap \bar{A} \cap \bar{A} \cap \ldots \cap \bar{A}) = \\ = P(A) \cdot P(A) \cdot P(A) \cdot \ldots \cdot P(A) \cdot P(\bar{A}) \cdot P(\bar{A}) \cdot \ldots \cdot P(\bar{A}) = \\ = p \cdot p \cdot p \cdot \ldots \cdot p \cdot q \cdot q \cdot q \cdot \ldots \cdot q = p^k \cdot q^{n-k} \end{gathered} \tag{5.12}$$

Deoarece trebuie să ținem cont de toate grupurile de k elemente de tip A (succese) pot fi alese din totalul celor n, indiferent de ordinea (poziția) acestora, avem un număr de C_n^k variante posibile.

Media distribuției binomiale este:

$$\mu = n \cdot p \tag{5.13}$$

Varianța (dispersia) distribuției binomiale:

$$\sigma^2 = n \cdot p \cdot q = n \cdot p \cdot (1 - p) \tag{5.14}$$

Distribuția binomială este biparametrică, adică este caracterizată de doi parametri n și p.

Distribuția binomială este asimetrică în următoarele situații:

- Dacă $p < 0{,}5$, distribuția prezintă *asimetrie la dreapta;*
- Dacă $p > 0{,}5$, distribuția prezintă *asimetrie la stânga*;
- Dacă volumul eșantionului n este redus, asimetria este mai accentuată. Pe măsura creșterii lui n, distribuția devine mai simetrică și tinde spre distribuția normală (conform teoremei limită centrale).

Pentru a exemplifica, presupunem că avem un test grilă și vom construi distribuția probabilistică a numărului de răspunsuri corecte, în cazul în care pentru fiecare întrebare avem sugerate 4 variante posibile de răspuns, din care doar una este corectă ($p = 0{,}25$). Ne situăm în situația în care studentul nu și-a însușit nicio noțiune din curs și va încerca în mod aleatoriu să bifeze răspunsul corect pentru fiecare întrebare, sperând ca va avea o zi norocoasă. Vom analiza situația unor teste cu 10, 20 și 50 de întrebări, evaluând de fiecare dată șansele de a promova examenul, numărul mediu (așteptat) de întrebări la care va fi bifat răspunsul corect și cu cât se abat în medie răspunsurile corecte în jurul acestui nivel mediu.

a) $n = 10$ întrebări și $p = 0{,}25$ (un răspuns corect din 4 variante posibile)

Scanned with OKEN Scanner

Pentru început, vom construi sub formă tabelară (tabelul 5.3) distribuția probabilistică a numărului de răspunsuri corecte (toate variantele posibile de la 0 la 10), calculând probabilitățile asociate prin utilizarea formulei 5.11:

$$P(X=0) = C_{10}^{2} \cdot (0{,}25)^{0} \cdot (1-0{,}25)^{10-0} = 0{,}056314$$
$$P(X=1) = C_{10}^{1} \cdot (0{,}25)^{1} \cdot (1-0{,}25)^{10-1} = 0{,}187712$$
$$P(X=2) = C_{10}^{2} \cdot (0{,}25)^{2} \cdot (1-0{,}25)^{10-2} = 0{,}281568$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$P(X=10) = C_{10}^{10} \cdot (0{,}25)^{10} \cdot (1-0{,}25)^{10-10} = 0{,}000001$$

Tabelul 5.3. Distribuția probabilistică a numărului de răspunsuri corecte la un test grilă de 10 întrebări ($p = 0{,}25$ și $n = 10$)

Numărul de răspunsuri corecte (x_i)	Probabilitatea corespunzătoare $P(x_i)$
0	0,056314
1	0,187712
2	0,281568
3	0,250282
4	0,145998
5	0,058399
6	0,016222
7	0,003090
8	0,000386
9	0,000029
10	0,000001

În continuare vom realiza o reprezentare grafică a datelor din tabelul 5.3 pentru a evidenția mai clar forma distribuției binomiale în acest caz. (figura 5.6).

Analizând figura 5.6, rezultă imediat faptul că, la testul cu 10 întrebări, având o singură variantă de răspuns corect din 4 posibile, avem o probabilitate maximă pentru studentul total nepregătit să răspundă corect la 2 întrebări $P(X=2) = 0.281568$, urmată de $P(X=3) = 0.250282$. Dacă cumulăm aceste 2 variante $P(X=2) + P(X=3) = 0{,}531850$ obținem o probabilitate totală de peste 53% ca studentul corect să răspundă la 2 sau la 3 întrebări, fără însă ca acest lucru să-i asigure promovarea examenului. De asemenea, probabilitatea ca studentul total nepregătit pentru examen să obțină în cadrul acestui test nota 10 este 0,000001.

Scanned with OKEN Scanner

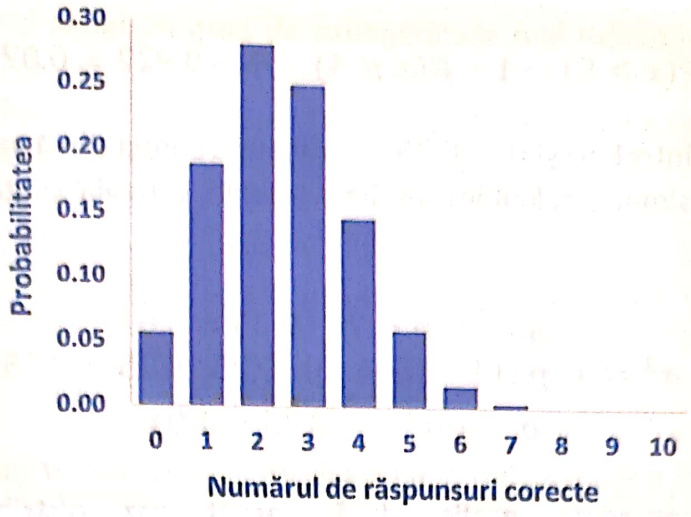


Figura 5.6. Distribuția probabilistică a răspunsurilor corecte ($p = 0{,}25$ și $n = 10$)

În continuare, calculăm media, varianța și deviația standard a acestei distribuții, utilizând formulele (5.13) și (5.14):

$$\mu = n \cdot p = 10 \cdot 0{,}25 = 2{,}50$$
$$\sigma^2 = n \cdot p \cdot (1-p) = 10 \cdot 0{,}25 \cdot 0{,}75 = 1{,}88$$
$$\sigma = \sqrt{\sigma^2} = \sqrt{1{,}88} = 1{,}37$$

Rezultă imediat faptul că, în cazul unui test grilă cu 10 întrebări, un student total nepregătit va putea spera să răspundă corect în medie la 2,5 întrebări cu o deviație standard σ de $\pm 1{,}37$ întrebări (sau calculând procentual, prin raportare la medie $\pm$ 54,8%). Media și deviația standard se pot calcula și utilizând formulele prezentate anterior (5.7) și (5.8), care sunt valabile în cazul tuturor distribuțiilor discrete.

Pentru a determina probabilitatea ca studentul nepregătit să reușească să obțină 5 sau mai mult răspunsuri corecte la testul grilă cu 10 întrebări anterior și astfel să promoveze examenul vom face următorul calcul:

$$P(x \geq 5) = P(5) + P(6) + P(7) + P(8) + P(9) + P(10) = 0{,}078 \text{ sau } 7{,}8\%$$

În concluzie, șansele unui student care este total nepregătit să promoveze un examen cu 10 întrebări grilă, având fiecare un singur răspuns corect din 4 posibile, sunt de aproximativ 7,8%.

Având în vedere că promovarea și nepromovarea examenului reprezintă evenimente complementare, a căror probabilitate cumulată însumează valoarea 1 (evenimentul sigur), putem determina șansele de promovare și ca diferență între 1 și probabilitatea de a nu promova acest examen, rezultatul fiind identic:

Scanned with OKEN Scanner

$$P(x \geq 5) = 1 - P(x \leq 4) = 1 - 0.922 = 0{,}078$$

b) $n = 20$ întrebări și $p = 0{,}25$ (un răspuns corect din 4 variante posibile)

În mod similar, calculăm media, varianța și deviația standard a acestei distribuții:

$$\mu = n \cdot p = 20 \cdot 0{,}25 = 5{,}0$$
$$\sigma^2 = n \cdot p \cdot (1-p) = 20 \cdot 0{,}25 \cdot 0{,}75 = 3{,}75$$
$$\sigma = \sqrt{\sigma^2} = \sqrt{3{,}75} = 1{,}94$$

Vom reprezenta grafic și în acest caz distribuția binomială corespunzătoare (figura 5.7):

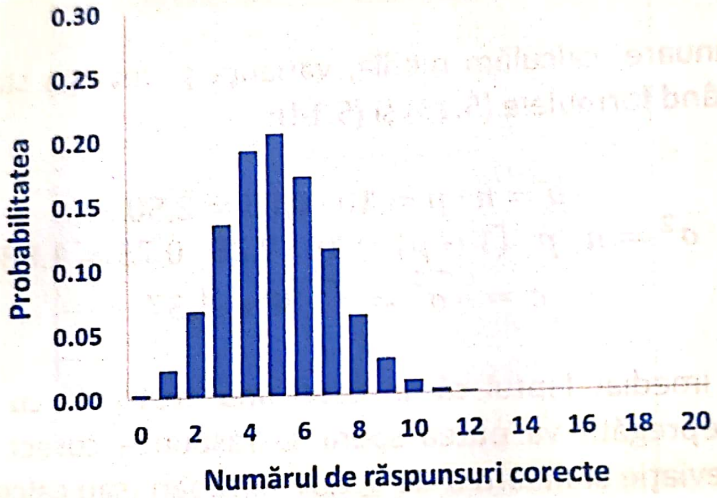


Figura 5.7. Distribuția probabilistică a răspunsurilor corecte ($p = 0{,}25$ și $n = 20$)

În acest caz, un student total nepregătit va putea spera să răspundă corect în medie la 5 întrebări cu o deviație standard σ de $\pm 1{,}94$ întrebări ($\pm$ 38,8%).

Probabilitatea de a promova examenul este:

$$P(x \geq 10) = P(10) + P(11) + P(12) + \cdots + P(20) = 0{,}014 \text{ sau } 1{,}4\%$$

Observăm o scădere importantă a probabilității de promovare a examenului, în situația când numărul de întrebări a crescut de la 10 la 20 (de la 7,8% la doar 1,4%). Cu alte cuvinte „norocul" începe să joace un rol mult mai redus, fiind astfel evidențiat mult mai clar nivelul real de pregătire a studentului. De asemenea și nivelul variabilității s-a diminuat, de la 54,8% până la 38,8%,

Scanned with OKEN Scanner

datele prezentând astfel un grad de omogenitate mai ridicat, care a condus la creșterea puterii de reprezentativitate a mediei.

c) $n = 50$ întrebări și $p = 0{,}25$ (un răspuns corect din 4 variante posibile)
Determinăm media, varianța și deviația standard:

$$\mu = n \cdot p = 50 \cdot 0{,}25 = 12{,}50$$

$$\sigma^2 = n \cdot p \cdot (1 - p) = 50 \cdot 0{,}25 \cdot 0{,}75 = 9{,}38$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{9{,}38} = 3{,}06$$

Deoarece este foarte utilă pentru înțelegerea în profunzime a fenomenului, vom realiza și în acest caz reprezentarea grafică a distribuției (figura 5.8):

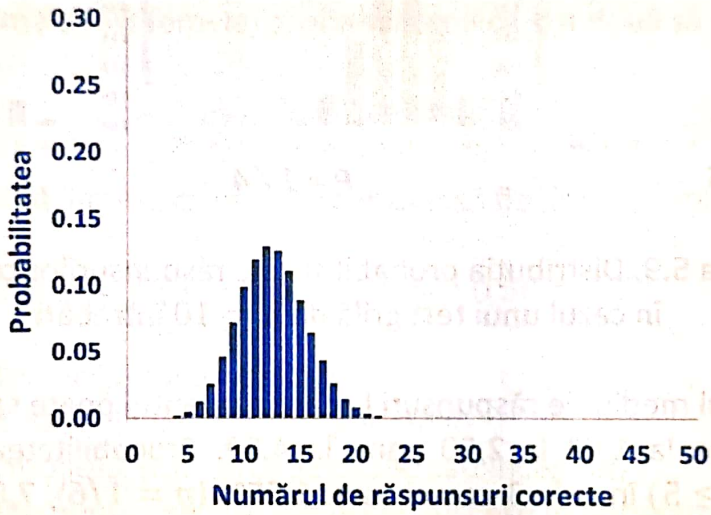


Figura 5.8. Distribuția probabilistică a răspunsurilor corecte
($p = 0{,}25$ și $n = 50$)

În această situație, probabilitatea de promovare (studentul să răspundă corect la cel puțin 25 de întrebări din totalul de 50) se apropie extrem de mult de valoarea zero (evenimentul imposibil):

$$P(x \geq 25) = P(25) + P(26) + P(27) + \cdots + P(50) = 0{,}0001$$

Prin urmare, este „aproape" imposibil să promovezi un examen grilă cu 50 de întrebări, având 4 variante de răspuns, dintre care doar una corectă, dacă ești total nepregătit la disciplina supusă evaluării. Nivelul variabilității numărului de răspunsuri corecte ale studenților în jurul nivelului mediu s-a redus și mai mult, comparativ cu situațiile anterioare, până la o valoare $\pm 24{,}5\%$. Media determinată în acest al treilea caz $\mu = 12{,}50$ ne arată că studentul nepregătit va răspunde corect cel mai probabil la aproximativ un sfert din totalul întrebărilor

Scanned with OKEN Scanner

(la fel ca și în situațiile anterioare). Cu toate acestea, datorită unui nivel de variabilitate (împrăștiere) mult mai redus la testul cu 50 de întrebări, șansa lui de a promova examenul datorată exclusiv „norocului" este aproape nulă.

Un alt aspect esențial se referă la faptul că, pe măsura creșterii volumului eșantionului (numărul de întrebări din test de la 10 la 50), distribuția binomială a devenit din ce în ce mai simetrică în jurul valorii medii, apropiindu-se foarte mult de alura clopotului lui Gauss (curba normală).

Interesant este să analizăm și situația examenelor cu 10 întrebări grilă, care pot avea 1 răspuns corect din șase, din patru sau din două posibile (figura 5.9).

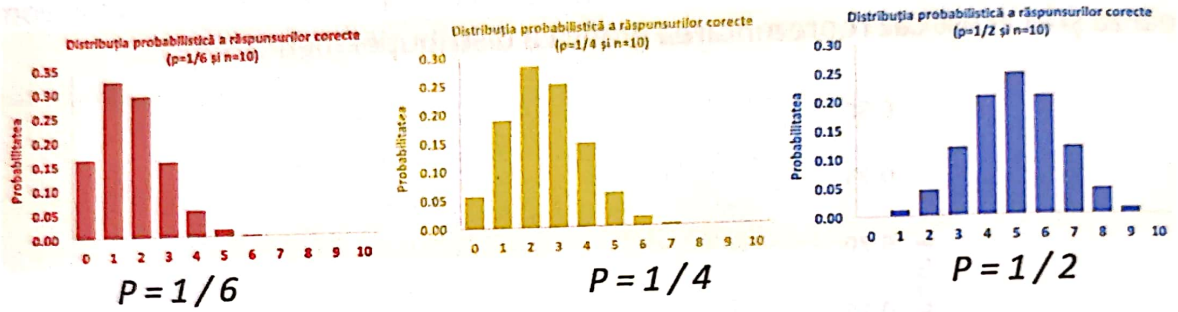

Distribuția probabilistică a răspunsurilor corecte (p=1/6 și n=10)
Probabilitatea
P = 1 / 6
Distribuția probabilistică a răspunsurilor corecte (p=1/4 și n=10)
Probabilitatea
P = 1 / 4
Distribuția probabilistică a răspunsurilor corecte (p=1/2 și n=10)
Probabilitatea
P = 1 / 2

Figura 5.9. Distribuția probabilistică a răspunsurilor corecte în cazul unui test grilă de $n = 10$ întrebări

Numărul mediu de răspunsuri la care studentul poate spera că răspunde corect crește de la 1,67 la 2,50 până la 4,99. Probabilitatea de a promova examenul $P(x \geq 5)$ în cele 3 cazuri este: 1,55% ($p = 1/6$), 7,81% ($p = 1/4$) și 62,30% ($p = 1/2$). Observăm și diminuarea treptată a asimetriei la dreapta a distribuției probabilistice a numărului de răspunsuri corecte, pe măsură ce numărul de variante de răspuns scade și implicit probabilitatea de a răspunde corect datorită „norocului" crește. Astfel, în situația unui test grilă de 10 întrebări care au doar două variante de răspuns, din care una corectă, probabilitatea de a promova examenul este de peste 40 de ori mai mare, comparativ cu un test grilă la care fiecare întrebare are șase variante de răspuns, din care una corectă.

2. DISTRIBUȚIA HIPERGEOMETRICĂ

Distribuția hipergeometrică este strâns legată de distribuția binomială, evidențiind probabilitatea obținerii a p succese din n probe (încercări) când există două rezultate posibile (succes sau insucces) pentru fiecare probă a experimentului. Diferența esențială între cele două distribuții probabilistice însă este faptul că, în cazul distribuției hipergeometrice probele nu sunt independente, probabilitatea obținerii unui succes modificându-se de la o încercare la alta. Din acest considerent, una dintre cele mai importante aplicații ale distribuției probabilistice hipergeometrice se referă la eșantionarea fără

Scanned with OKEN Scanner

revenire din cadrul unei populații finite, în care o unitate, care a fost selectată în vederea includerii în eșantion, nu este reintrodusă în populația de referință, aspect care conduce în mod inevitabil la diminuarea probabilității de selecție ulterioară a unui element din acea categorie.

Pentru a exemplifica, considerăm că avem o echipă de 5 cercetători, compusă din trei femei și doi bărbați. Dintre aceștia, doar doi membrii trebuie să reprezinte echipa la un Congres. S-a decis să fie aleși în mod aleatoriu 2 membri în vederea participării. Care este probabilitatea ca aceste două persoane să fie de sex feminin?

Probabilitatea ca prima persoană să fie femeie este:

$$P(F) = 3/5 = 0{,}60$$

Dacă prima a fost femeie, probabilitatea ca și a două să fie femeie este:

$$P_F(F) = 2/4 = 0{,}50$$

Deci probabilitatea ca cei doi membrii să fie de sex feminin este:

$$P = P(F) \cdot P_F(F) = 0{,}60 \cdot 0{,}50 = 0{,}30$$

Vom utiliza următoarele notații:
N = volumul populației (colectivitatea generală)
n = volumul eșantionului (numărul de probe sau încercări)
r = numărul de elemente considerate drept succese în populația generală
x = numărul de elemente considerate succese în eșantion

Funcția de probabilitate hipergeometrică reprezintă probabilitatea obținerii în cadrul unui eșantion de n unități, selectate fără revenire dintr-o populație de volum N în care avem un număr r de elemente considerate drept succese, a unui număr de x unități etichetate drept succese (posedă un anumit nivel al variabilei) și a unui număr de $n - x$ unități etichetate ca insuccese:

$$P(X = x) = \frac{C_r^x \cdot C_{N-r}^{n-x}}{C_N^n} \tag{5.15}$$

Media distribuției hipergeometrice este:

$$\mu = n \cdot \frac{r}{N} \tag{5.16}$$

Varianța (dispersia) distribuției hipergeometrice:

Scanned with OKEN Scanner

$$\sigma^2 = n \cdot \frac{r}{N}\left(1 - \frac{r}{N}\right)\frac{N-r}{N-1} \tag{5.17}$$

Distribuția hipergeometrică depinde de trei parametri: N (volumul populației), n (volumul eșantionului) și r (numărul de succese din populația generală). În practică, o distribuție hipergeometrică poate fi de obicei aproximată printr-o distribuție binomială, întrucât în situația în care volumul eșantionului nu depășește 5% din cel al populației, există o diferență redusă între eșantionarea cu și fără revenire.

De exemplu, să presupunem faptul că la o universitate avem în anii I și II 2000 de studenți (1500 de sex feminin și 500 de sex masculin). Din această colectivitate vom extrage în mod aleatoriu două eșantioane formate din 10 și respectiv 50 de studenți. Vom evidenția distribuția probabilistică hipergeometrică a numărului de fete care intră în componența fiecărui eșantion, numărul mediu (așteptat) de persoane de sex feminin și mărimea medie a abaterilor nivelurilor individuale în jurul acestui nivel, pentru fiecare eșantion.

Eșantionarea va fi efectuată fără revenire, fiind „riscant" să apelăm la modelul binomial, deoarece ne putem situa în situația în care „cei 10 studenți aleși pot fi în fapt doar 2 persoane, una selectată de 6 ori, iar cealaltă de 4 ori".

a) $N = 2000$ studenți, $n = 10$ studenți, $r = 1500$ studente în populația generală, $x = 0, 1, 2, 3, \ldots, 10$ studente la nivelul eșantionului extras.

Construim sub formă tabelară (tabelul 5.4) distribuția probabilistică a numărului de persoane de sex feminin din eșantion (toate variantele posibile de la 0 la 10):

Tabelul 5.4. Distribuția probabilistică a numărului de persoane de sex feminin din eșantion

Numărul de fete din eșantion (x_i)	**Probabilitatea corespunzătoare $P(x_i)$**
0	0,000001
1	0,000027
2	0,000373
3	0,003024
4	0,016035
5	0,058155
6	0,146070
7	0,250910
8	0,282085
9	0,187429
10	0,055891

Scanned with OKEN Scanner

Probabilitățile asociate prezentate în tabelul 5.4. au fost determinate prin utilizarea formulei 5.15:

$$P(X=0)=\frac{C_{1500}^{0}\cdot C_{500}^{10-0}}{C_{2000}^{10}}=0{,}000001 \qquad P(X=1)=\frac{C_{1500}^{1}\cdot C_{500}^{10-1}}{C_{2000}^{10}}=0{,}000027$$

$$P(X=2)=\frac{C_{1500}^{2}\cdot C_{500}^{10-2}}{C_{2000}^{10}}=0{,}000373 \ldots\ldots P(X=10)=\frac{C_{1500}^{10}\cdot C_{500}^{10-10}}{C_{2000}^{10}}=0{,}055891$$

Vom realiza o reprezentare grafică a datelor din tabelul 5.4 pentru a evidenția mai clar forma distribuției hipergeometrice în acest caz. (figura 5.10).

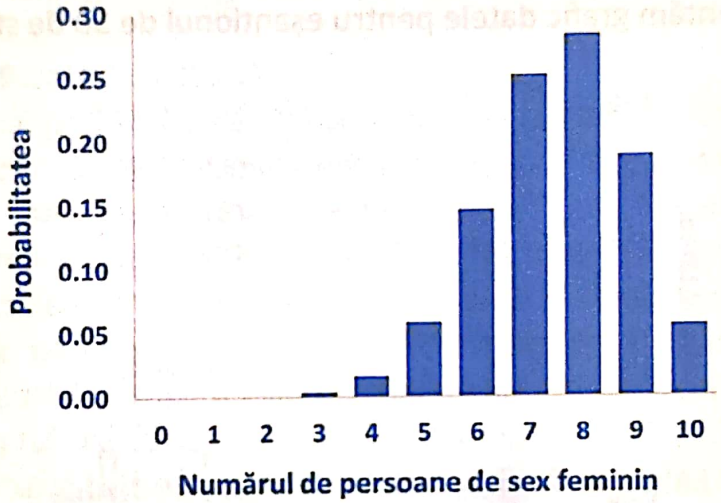


Figura 5.10. Distribuția hipergeometrică a persoanelor de sex feminin ($N=2000, n=10$ și $r=1500$)

Din analiza figurii 5.10 și a tabelului 5.4, se observă imediat faptul că în eșantionul de 10 studenți extras în mod aleatoriu dintr-o populație de 2000 de persoane, din care 1500 sunt de sex feminin, vom avea cu probabilitate maximă 8 fete $P(X=8)=0{,}282085$.

Calculăm media, varianța și deviația standard a acestei distribuții, utilizând formulele (5.16) și (5.17):

$$\mu=n\cdot\frac{r}{N}=10\cdot\frac{1500}{2000}=7{,}50$$

$$\sigma^2=n\cdot\frac{r}{N}\left(1-\frac{r}{N}\right)\frac{N-r}{N-1}=10\cdot\frac{1500}{2000}\left(1-\frac{1500}{2000}\right)\frac{2000-1500}{2000-1}=0{,}47$$

$$\sigma=\sqrt{\sigma^2}=\sqrt{0{,}47}=0{,}68$$

Scanned with OKEN Scanner

Prin urmare, în eșantionul de 10 studenți extrași în mod aleatoriu putem spera să avem un număr mediu de 7,50 persoane de sex feminin, cu o abatere medie σ de $\pm$ 0,68 persoane.

b) $N = 2000$ studenți, $n = 50$ studenți, $r = 1500$ studente în populația generală $x = 0, 1, 2, 3, \ldots, 50$ studente la nivelul eșantionului extras

În mod asemănător, putem determina probabilitatea de a avea $X = 38$ studente în eșantionul de 50 studenți este:

$$P(X = 38) = \frac{C_{1500}^{38} \cdot C_{500}^{12}}{C_{2000}^{50}} = 0{,}1309$$

Reprezentăm grafic datele pentru eșantionul de 50 de studenți:

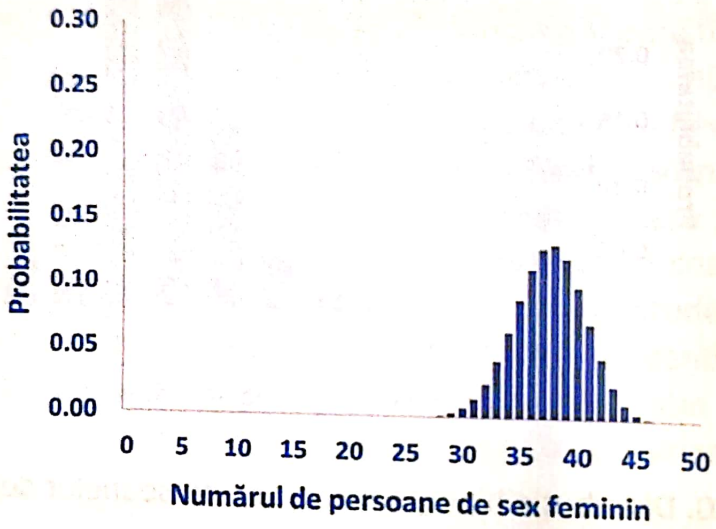


Figura 5.11. Distribuția hipergeometrică a persoanelor de sex feminin ($N = 2000, n = 50$ și $r = 1500$)

Determinăm și în acest caz media, varianța și deviația standard a acestei distribuții, utilizând formulele (5.16) și (5.17):

$$\mu = n \cdot \frac{r}{N} = 50 \cdot \frac{1500}{2000} = 37{,}50$$

$$\sigma^2 = n \cdot \frac{r}{N}\left(1 - \frac{r}{N}\right)\frac{N - r}{N - 1} = 50 \cdot \frac{1500}{2000}\left(1 - \frac{1500}{2000}\right)\frac{2000 - 1500}{2000 - 1} = 2{,}34$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{2{,}34} = 1{,}53$$

Scanned with OKEN Scanner

În eșantionul de 50 studenți extrași în mod aleatoriu din populația de 2000 de persoane putem spera să avem un număr mediu de 37,50 persoane de sex feminin, cu o abatere medie σ de $\pm$ 1,53 persoane.

Observăm faptul că, în situația în care utilizăm un eșantion de volum superior n=50 față de cel extras inițial ($n = 10$) asimetria moderată la stânga (negativă) se reduce foarte mult, iar distribuția hipergeometrică devine din ce în ce mai simetrică, apropiindu-se de alura clopotului lui Gauss. De asemenea, șansa ca în eșantionul extras dintr-o populație formată în proporție de 75% din persoane de sex feminin să fie mai puțin de 50% studente scade de la 7,76% în cazul eșantionului de 10 persoane la 0,01% pentru cel de 50 persoane (eșantionul este mai reprezentativ).

3. DISTRIBUȚIA POISSON

Distribuția probabilistică Poisson poartă numele matematicianului și fizicianului francez Siméon Denis Poisson (1781-1840), profesor la École Polytechnique și membru al Academiei Franceze de Științe, care a studiat-o în detaliu în lucrarea sa din 1837, intitulată *Recherches sur la probabilité des jugements en matière criminelle et en matière civile*. Reprezintă o distribuție discretă utilizată pentru a modela probabilitatea ca un anumit număr de evenimente să apară într-un interval de timp/spațiu bine definit (fix), atunci când se cunoaște faptul că evenimentele au loc independent într-un interval comparativ cu altul și cu o rată medie constantă. Probabilitatea unui eveniment este direct proporțională cu lungimea intervalului în care acesta are loc, iar numărul de apariții într-un anumit interval nu afectează pe cel din oricare altul.

Distribuția Poisson este o variantă limitată a distribuției binomiale, atunci când probabilitatea unui succes este foarte redusă și numărul de probe (volumul eșantionului) n este mare. Din acest considerent, se face frecvent referire la ea la fiind *„legea evenimentelor improbabile"*. Este utilizată pentru a descrie distribuția erorilor care apar la introducerea datelor, numărul de imperfecțiuni care apar la mașinile noi, numărul de componente defecte care apar în cadrul unui transport, numărul de pacienți internați în fiecare oră într-un spital, rata de apariție a unei boli rare, numărul de bagaje pierdute în cadrul unui zbor pentru o anumită companie aeriană, etc.

Funcția de probabilitate a distribuției Poisson este:

$$P(X = x) = \frac{\mu^x \cdot e^{-\mu}}{x!} \tag{5.18}$$

în care:

μ = numărul mediu de apariții (succese) dintr-un anumit interval

x = numărul de apariții (succese)

$P(X = x)$ = probabilitatea aferentă fiecărei valori particulare a lui x

Scanned with OKEN Scanner

$e = 2{,}71828$ (constanta lui Euler sau baza logaritmilor naturali)

Media distribuției Poisson este: $\mu = n \cdot p$ (5.19)

Varianța distribuției Poisson este $\sigma^2 = n \cdot p$ (5.20)

Dacă în cazul distribuției probabilistice binomiale exista un număr fix de probe, variabila aleatoare în cazul unei distribuții Poisson poate avea un număr infinit de valori: $0, 1, 2, 3, 4, 5, 6, 7, \ldots$

În orice caz, probabilitățile aferente acestor valori ale variabilei devin din ce în ce mai mici după primele apariții (succese).

De exemplu, o policlinică dorește să utilizeze distribuția Poisson în vederea estimării numărului posibil de disfuncționalități ale rețelei informatice care pot surveni într-o lună. Presupunând că este cunoscut anterior faptul că numărul mediu de probleme ale rețelei informatice este de 5 pe lună ($\mu = 5$), vom construi distribuția probabilistică a numărului de disfuncționalități ale rețelei care apar într-o lună. Variabila aleatoare discretă este numărul de disfuncționalități ale rețelei informatice (0,1,2,3,...). Intervalul (de timp) este luna calendaristică (probabilitatea apariției problemelor la rețeaua informatică a policlinicii crește cu cât numărul de zile lucrătoare dintr-o lună este mai mare).

Tabelul 5.5. Distribuția probabilistică a numărului de disfuncționalități ale rețelei informatice

Numărul de probleme ale rețelei x_i	Probabilitatea corespunzătoare $P(x_i)$
0	0,006738
1	0,033690
2	0,084224
3	0,140374
4	0,175467
5	0,175467
6	0,146223
7	0,104445
8	0,065278
9	0,036266
10	0,018133
11	0,008242
12	0,003434

Scanned with OKEN Scanner

Numărul de disfuncționalități ale rețelei dintr-o anumită lună este independent de cel apărut în celelalte luni. Distribuția probabilistică a acestei variabile este prezentată în tabelul 5.5.

Probabilitățile asociate numărului de disfuncționalități ale rețelei informatice din tabelul 5.5 au fost calculate prin utilizarea formulei (5.18), astfel:

$$P(X=0)=\frac{5^0\cdot e^{-5}}{0!}=0{,}006738 \qquad P(X=1)=\frac{5^1\cdot e^{-5}}{1!}=0{,}033690$$

$$P(X=2)=\frac{5^2\cdot e^{-5}}{2!}=0{,}084224 \quad \ldots\ldots \quad P(X=12)=\frac{5^{12}\cdot e^{-5}}{12!}=0{,}003434$$

În continuare, vom realiza o reprezentare grafică a datelor din tabelul 5.5, în scopul de a contura vizual forma distribuției Poisson în cazul exemplului nostru. (figura 5.12).

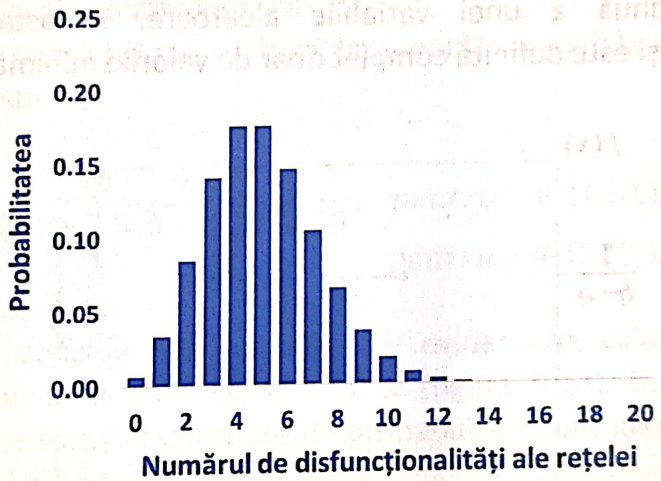


Figura 5.12. Distribuția Poisson a numărului de disfuncționalități ale rețelei informatice ($\mu = 5$)

Ca și distribuțiile binomială și hipergeometrică, distribuția Poisson aparține unei familii de distribuții discrete. Pentru a construi distribuția Poisson este necesar doar un singur parametru – media μ a variabilei aleatoare (numărul de defecte, erorile, etc.).

Distribuția Poisson prezintă un anumit grad de asimetrie la dreapta, care este mai accentuat în special în situațiile când parametrul $\mu < 5$. Pe măsură ce crește parametrul μ (media variabilei dintr-un anumit interval), distribuția Poisson devine din ce în ce mai simetrică și poate fi aproximată printr-o distribuție normală, cu media și varianța egale cu μ.

Scanned with OKEN Scanner

5.6. DISTRIBUȚII PROBABILISTICE CONTINUE – UNIFORMĂ ȘI NORMALĂ

După cum am văzut anterior, în cazul distribuțiilor aleatoare discrete se poate determina probabilitatea ca variabila analizată sa ia o anumită valoare particulară. În situația distribuțiilor probabilistice continue însă, o variabilă aleatoare continuă poate avea orice valoare în cadrul unui interval. Deoarece există un număr infinit de valori în oricare interval, nu mai are sens să discutăm despre probabilitatea ca variabila aleatoare continuă să ia o valoare exactă, ci de probabilitatea ca aceasta să fie situată în cadrul limitelor unui interval.

1. DISTRIBUȚIA UNIFORMĂ

Distribuția probabilistică uniformă reprezintă varianta cea mai simplă de distribuție continuă a unei variabile aleatoare. Aceasta are o formă dreptunghiulară și este definită complet doar de valorile minimă și maximă ale ei (figura 5.13).

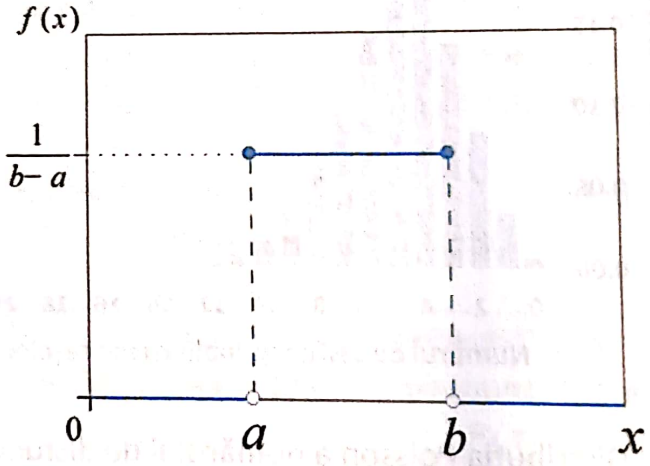


Figura 5.13. Distribuția uniformă a variabilei continue X

Funcția *densității de probabilitate* a distribuției uniforme este:

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & pentru\ x \in [a,b] \\ 0, & pentru\ x \notin [a,b] \end{cases} \qquad (5.21)$$

Media valorilor unei variabile uniform distribuite în intervalul (a, b):

$$\mu = \frac{a+b}{2} \qquad (5.22)$$

Scanned with OKEN Scanner

Varianța (dispersia) distribuției uniforme continue:

$$\sigma^2 = \frac{(b-a)^2}{12} \tag{5.23}$$

De exemplu, considerăm o variabilă X care reprezintă timpul total de zbor al unui avion între două orașe. Presupunem faptul că timpul de zbor poate înregistra orice valoare între 120 și 140 de minute. Deoarece variabila aleatoare X poate lua orice valoare în interiorul intervalului 120 - 140 minute, aceasta poate fi considerată drept o variabilă continuă. Considerăm că avem suficiente informații pentru a afirma probabilitatea ca un zbor să dureze între 120 și 121 de minute este egală cu probabilitatea oricărui alt zbor de a se situa în orice alt interval de 1 minut pe toată plaja de valori de la 121 la 140 minute inclusiv. Astfel, cu toate intervalele de 1 minut egal probabile, putem afirma faptul că variabila X urmează o distribuție continuă uniformă.

Funcția densității de probabilitate pentru variabila aleatoare „timpul total de zbor" este:

$$f(x) = \begin{cases} \dfrac{1}{140-120} = \dfrac{1}{20}, & pentru\ x \in [120,140] \\ 0, & pentru\ x \notin [120,140] \end{cases}$$

Graficul densității de probabilitate evidențiază valorile funcției $f(x)$ pentru fiecare valoare particulară a lui x (figura 5.14). În cazul distribuției uniforme, valorile funcției densitate de probabilitate sunt identice pentru toate valorile variabilei x.

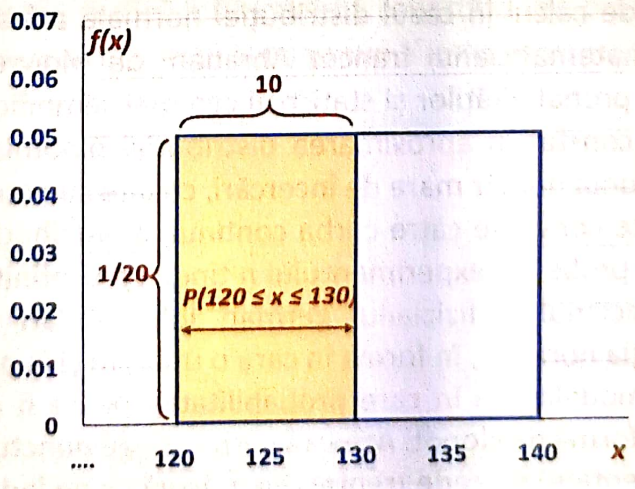


Figura 5.14. Distribuția uniformă a variabilei continue „timpul total de zbor"

Scanned with OKEN Scanner

În mod diferit față de cazul distribuțiilor discrete, în situația distribuțiilor continue funcția densității de probabilitate nu reprezintă probabilități, ci „înălțimea funcției" pentru fiecare valoare particulară lui x. Pentru o variabilă aleatoare continuă, probabilitatea reprezintă șansa ca variabila să ia o valoare situată într-un anumit interval.

Media sau valoarea tipică a timpului total de zbor exprimat în minute:

$$\mu = \frac{120 + 140}{2} = 130$$

Varianța variabilei:

$$\sigma^2 = \frac{(140 - 120)^2}{12} = 33{,}33$$

Deviația standard a timpilor de zbor în jurul valorii medii:

$$\sigma = \sqrt{33{,}33} = 5{,}77 \ minute$$

2. DISTRIBUȚIA NORMALĂ

Distribuția normală, cunoscută și sub denumirea de distribuție Gausiană, este probabil cea mai importantă distribuție probabilistică utilizată pentru a descrie o variabilă aleatoare continuă, având multiple aplicații practice (greutatea, înălțimea, scorurile IQ, diferite măsurători științifice, cantitatea de precipitații, notele la un examen, etc.). Pentru a utiliza distribuția probabilistică normală, variabila aleatoare analizată trebuie să fie continuă.

Formula de calcul în cazul distribuției normale a fost publicată pentru prima dată de matematicianul francez Abraham de Moivre în anul 1733. În domeniul teoriei probabilităților și statisticii cea mai semnificativă contribuție a lui de Moivre a constat în aproximarea distribuției binomiale prin distribuția normală în cazul unui număr mare de încercări, cu alte cuvinte forma distribuției binomiale discrete converge către curba continuă a distribuției normale, atunci când numărul de probe ale experimentului n tinde spre infinit.

Matematicianul și fizicianul german Johann Carl Friedrich Gauss a introdus distribuția normală, în forma în care o utilizăm și în prezent, prezentând în mod explicit modalitatea în care probabilitatea poate fi reprezentată grafic printr-o curbă în formă de clopot. Această curbă atinge punctul de maxim în jurul valorii medii (așteptate) și scade treptat, pe măsură ce ne îndepărtăm de nivelul mediu, în direcția celor două limite ($-\infty$ și $+\infty$).

Funcția Gauss-Laplace este caracterizată de doi parametri: media (μ), și dispersia (σ^2):

$$N(\mu, \sigma^2)$$

Scanned with OKEN Scanner

Expresia *densității de probabilitate* este dată de formula:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-\mu)^2}{2\cdot\sigma^2}}$$

unde X este o variabilă continuă: $-\infty < x < +\infty$ (5.24)

Funcția densitate de probabilitate poate fi reprezentă sub formă grafică:

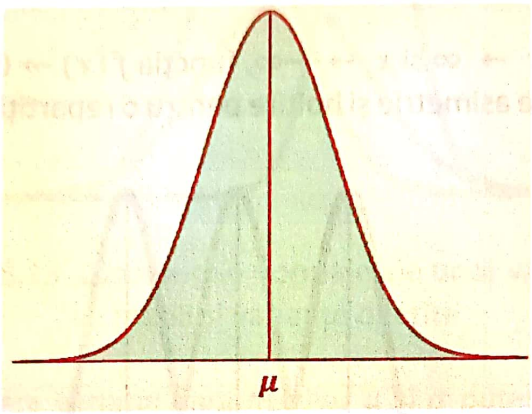


Figura 5.15. Distribuția normală a unei variabile continue

Funcția de repartiție se calculează astfel:

$$F(x) = \int_{-\infty}^{x} f(x)dx = \frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^{x} e^{\frac{-(x-\mu)^2}{2\cdot\sigma^2}} dx \quad (5.25)$$

Reprezentarea grafică a funcției de repartiție (de probabilitate) în cazul distribuției normale a unei variabile continue X este prezentată în figura 5.16:

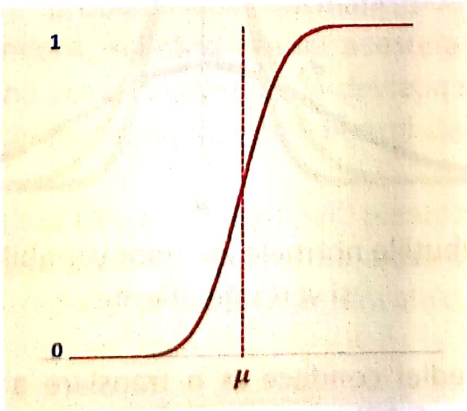


Figura 5.16. Funcția de repartiției a unei variabile continue

Scanned with OKEN Scanner

Proprietățile repartiției normale:

- admite un singur punct de maxim ($x = \mu$), fiind unimodală
- este simetrică in raport cu dreapta $x = \mu$
- densitatea de probabilitate are formă de clopot → „clopotul lui Gauss"
- media aritmetică, mediana și modul coincid
- suprafața totală cuprinsă între curba și axa absciselor este 1, reprezentând probabilitatea evenimentului sigur
- punctele de pe abscisă, de valoare $\mu - \sigma$, $\mu + \sigma$, sunt puncte de inflexiune
- în punctele $x \to \infty$ si $x \to -\infty$, funcția $f(x) \to 0$
- coeficienții de asimetrie și boltire pentru o repartiție normală au valoarea zero

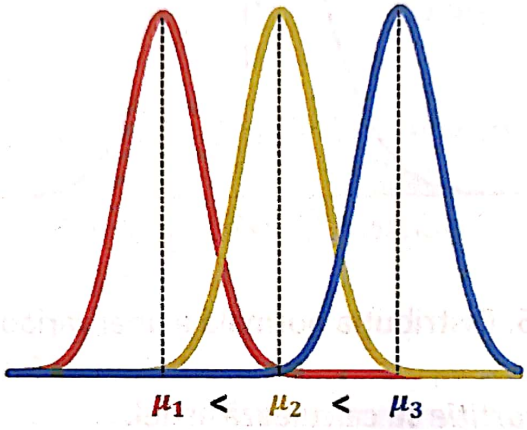


Figura 5.17. Distribuțiile normale ale unor variabile cu medii diferite și aceeași varianță

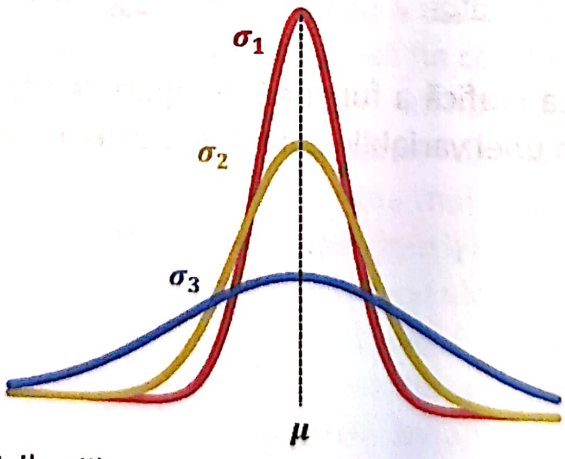


Figura 5.18. Distribuțiile normale ale unor variabile cu aceeași medie și varianțe diferite

- modificarea mediei conduce la o translare a curbei de-a lungul axei absciselor (figura 5.17)
- curba densității de probabilitate este cu atât mai boltită cu cât valoarea dispersiei este mai mică (figura 5.18)

Scanned with OKEN Scanner

Datorită acestor caracteristici, putem afirma faptul că media μ reprezintă un parametru de localizare și deviația standard σ un parametru al formei distribuției normale.

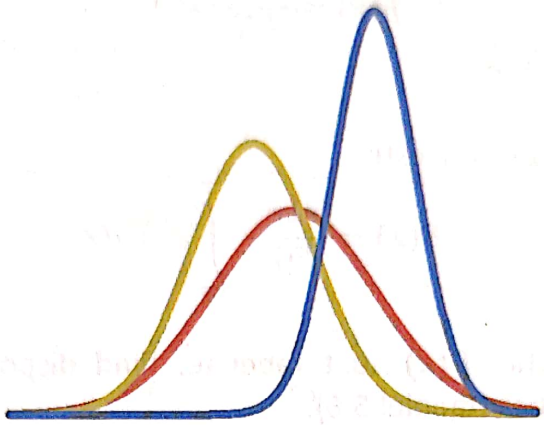


Figura 5.19. Distribuțiile normale ale unor variabile cu medii și varianțe diferite

Având în vedere variația parametrilor μ și σ putem vorbi de un număr practic nelimitat de repartiții normale, fiecare având media și deviația sa standard particulare (figura 5.19).

5.7. DISTRIBUȚIA NORMALĂ STANDARD

Caracteristicile prezentate anterior arată faptul că, distribuția normală este *„o familie de distribuții"*, în care fiecare membru se deosebește de altul în baza valorilor μ și σ. Cel mai important membru al acestei familii este distribuția normală standard $N(0,1)$, care are o medie egală cu zero ($\mu = 0$) și o deviație standard egală cu 1 ($\sigma = 1$).

Orice distribuție probabilistică normală poate fi convertită într-o distribuție normală standard, scăzând media acesteia (μ) din fiecare valoare observată (x_i) și împărțind această diferență la deviația standard (σ). Rezultatele obținute prin această schimbare de variabilă poartă denumirea de valori z sau *scoruri* z.

Astfel, presupunând că avem o variabilă aleatoare distribuită normal X, valorile z_i asociate fiecărui nivel x_i al lui X ne arată distanța acestuia față de media aritmetică μ a variabilei exprimată în unități ale deviației standard σ:

$$z = \frac{x_i - \mu}{\sigma} \tag{5.26}$$

Scanned with OKEN Scanner

Densitatea de probabilitate a distribuției normale standard se determină după formula:

$$f(z) = \frac{1}{\sqrt{2\pi}} \cdot e^{\frac{-z^2}{2}} \tag{5.27}$$

unde: $-\infty < z < +\infty$

Funcția de repartiție este:

$$F(z) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{z} e^{\frac{-z^2}{2}} dz \tag{5.28}$$

Valorile funcției $F(z)$ sunt tabelate, fiind disponibile în majoritatea manualelor de statistică (tabelul 5.6).

Tabelul 5.6. Valorile tabelate ale distribuției normale standard

DISTRIBUTIA NORMALA STANDARD

z	Probabilitati cumulate					
	0.00	0.01	0.02	0.03	0.04	0.05
0.00	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199
0.10	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596
0.20	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987
0.30	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368
...	...	...	...	...	...	...
1.40	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265
1.50	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394
1.60	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505
1.70	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599

Pentru a citi o valoare din tabel, localizăm, de exemplu, $z = 1{,}50$ în tabelul 5.6 și preluăm nivelul corespunzător al *probabilității cumulate* de 0,9332. Valoarea de 93,32% reprezintă probabilitatea ca o un anumit z extras în mod aleatoriu din populația Z să aibă o valoare cuprinsă între $-\infty$ și 1,50. Aceasta poate fi interpretată și ca proporția (frecvența relativă de apariție) a valorilor lui z cuprinse între -∞ și 1,50. Pentru $z = 1{,}65$, de exemplu, citim valoarea situată la intersecția liniei 1,60 și coloanei 0,05 și preluăm nivelul corespunzător al *probabilității cumulate* de 0,9505 sau 95,05%.

Probabilitatea ca o anumită variabilă de interes X, având media μ și deviația standard σ, să ia valori într-un interval (x_1, x_2) se determină, utilizând schimbarea de variabilă z:

$$P(x_1 < x < x_2) = P\left(\frac{x_1 - \mu}{\sigma} < z < \frac{x_2 - \mu}{\sigma}\right) =$$

Scanned with OKEN Scanner

$$= P\left(-\infty < z < \frac{x_2 - \mu}{\sigma}\right) - P\left(-\infty < z < \frac{x_1 - \mu}{\sigma}\right) \quad (5.29)$$

Exemplul 1: într-un lot de pacienți normal distribuit, având media $\mu = 90\ kg$ și abaterea medie pătratică (deviația standard) $\sigma = 15\ kg$, dorim să aflăm probabilitatea ca o persoană să aibă sub $65\ kg$.

Realizăm conversia distribuției normale de medie $\mu = 90\ kg$ și deviație standard $\sigma = 15\ kg$ într-o distribuție normală standard, de medie 0 și deviație standard 1. În acest scop, calculăm scorul z_i al valorii $x_i = 65\ kg$, utilizând formula (5.26):

$$z_i = \frac{x_i - \mu}{\sigma} = \frac{65 - 90}{15} = -1{,}667$$

Ulterior, vom determina probabilitatea ca valoarea z_i să fie mai mică decât $-1{,}667$, conform tabelului distribuției normale standard. Având în vedere faptul că distribuția normală standard este simetrică, putem calcula probabilitatea ca valoarea z_i să fie mai mare decât 1,667, aceasta fiind identică cu probabilitatea căutată $P(z_i < -1{,}667)$. Evenimentele $P(z_i < -1{,}667)$ și $P(z_i > 1{,}667)$ sunt complementare, astfel încât suma probabilităților lor de realizare este egală cu 1:

$$P(z_i > 1{,}667) + P(z_i < -1{,}667) = 1$$

Prin urmare:

$$P(z_i > 1{,}667) = 1 - P(z_i < -1{,}667) = 1 - 0{,}952 = 0{,}048$$

În situația în care nu dispunem de un tabel al distribuției normale standard tipărit, putem utiliza programul Excel, pentru a calcula direct probabilitatea $P(z_i < -1{,}667)$, utilizând formula:

$$= NORM.S.DIST(-1.667, Cumulativ = TRUE)$$

Aceasta va returna valoarea 0,048, rezultat similar cu cel obținut anterior prin utilizarea tabelului distribuției normale standard.

Prin urmare, probabilitatea ca o persoană care a fost extrasă în mod aleatoriu dintr-un lot de pacienți normal distribuit, de medie $\mu = 90\ kg$ și deviație standard $\sigma = 15\ kg$, să aibă sub $65\ kg$ este de doar 4,8%:

$$P(x_i < 65) = P\left(z_i < \frac{65 - 90}{15}\right) = P(z_i < -1{,}667) = 0{,}048$$

Scanned with OKEN Scanner

Se poate realiza și o reprezentare grafică (figura 5.20) pentru a facilita interpretarea rezultatului:

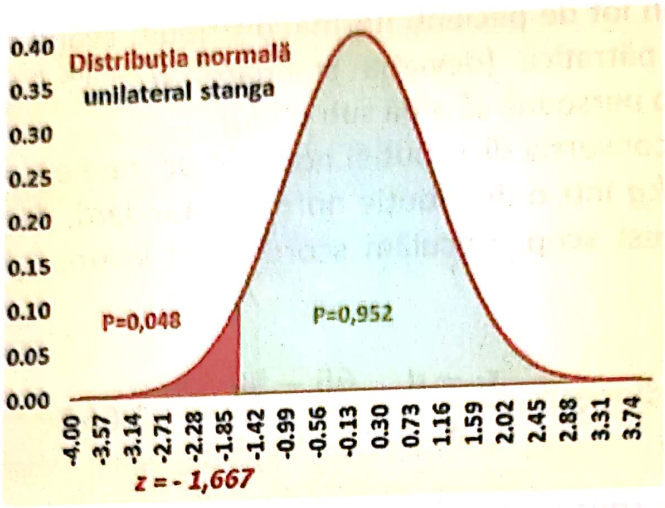


Figura 5.20 Probabilitatea ca un pacient să aibă sub 65 kg transpusă în distribuția normală standard

Exemplul 2: în cadrul aceluiași eșantion de pacienți, dorim să calculăm probabilitatea ca o persoană să aibă peste 100 kg.

Deoarece în tabelul distribuției normale standard avem probabilitățile cumulate ca scorul z_i să fie sub un anumit nivel dorit, iar noi dorim să determinăm această valoare pentru un nivel prag specificat (100 kg), vom apela la regula complementară, scăzând din 1 valoarea probabilității evenimentului contrar (opus):

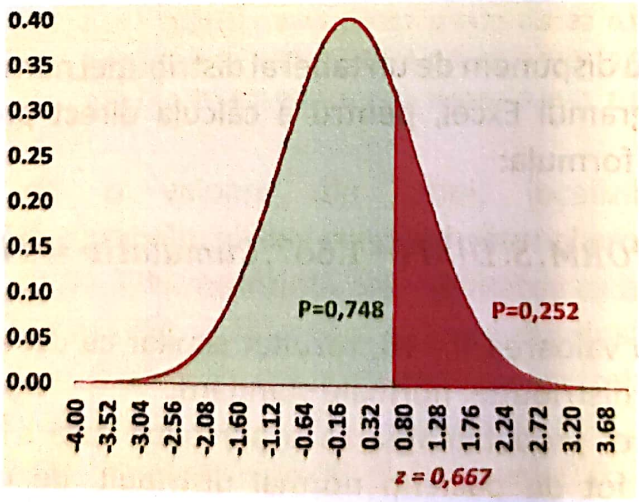


Figura 5.21 Probabilitatea ca un pacient să aibă peste 100 kg transpusă în distribuția normală standard

$$P(x_i > 100) = 1 - P(x_i < 100) = 1 - P\left(z_i < \frac{100 - 90}{15}\right) =$$

Scanned with OKEN Scanner

$$= 1 - P(z_i < 0{,}667) = 1 - P(-\infty < z_i < 0{,}667) = 1 - 0{,}748 = 0{,}252$$

Prin urmare, probabilitatea ca o persoană din acest lot de pacienți normal distribuit, de medie 90 kg și deviație standard 15 kg, să aibă peste 100 kg este destul de mare, respectiv de 25,2% (zona marcata cu roșu în figura 5.21).

Exemplul 3: În cadrul unei populații, nivelul de inteligență urmează o distribuție normală, cu un nivel mediu $\mu = 100$ și o deviație standard de $\sigma = 15$. Să se determine șansa să întâlnim o persoană cu IQ-ul cuprins între 120 și 130.

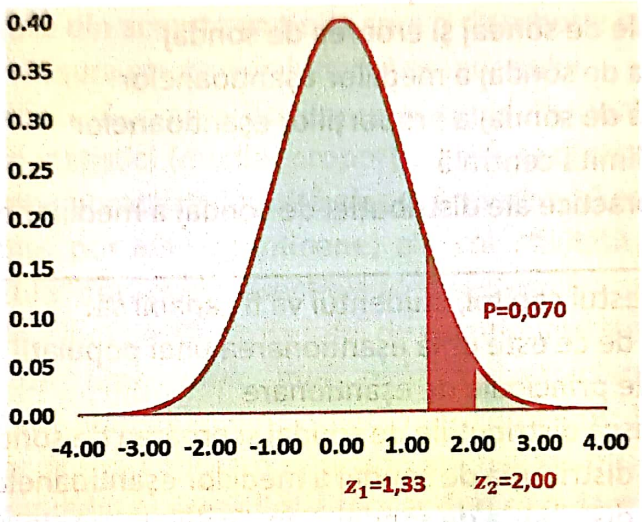


Figura 5.22 Probabilitatea ca o persoană să aibă IQ-ul cuprins între 120 și 130 transpusă în distribuția normală standard

Aplicăm relația 5.29 și obținem:

$$P(120 < x_i < 130) = P\left(\frac{120 - 100}{15} < z_i < \frac{130 - 100}{15}\right) =$$

$$= P(1{,}33 < z_i < 2{,}00) = P(-\infty < z_i < 2{,}00) - P(-\infty < z_i < 1{,}33) =$$

$$= 0{,}98 - 0{,}91 = 0{,}07$$

Astfel, în cadrul populației studiate care urmează o distribuție aproximativ normală, de medie 100 și deviație standard 15, probabilitatea de a întâlni o persoană cu scorul IQ cuprins în intervalul 120 – 130 este destul de redusă, aceasta înregistrând o valoare de 7% (aria marcată cu roșu în figura 5.22).

Scanned with OKEN Scanner

Capitolul. 6

METODE DE EȘANTIONARE ȘI TEOREMA LIMITĂ CENTRALĂ

SUBIECTE:

6.1. Avantajele eșantionării
6.2. Principalele metode de eșantionare
6.3. Distribuțiile de sondaj și eroarea de sondaj
6.4. Distribuția de sondaj a mediilor eșantioanelor
6.5. Distribuția de sondaj a proporțiilor eșantioanelor
6.6. Teorema limită centrală
6.7. Aplicații practice ale distribuției de sondaj a mediilor eșantioanelor

La finalizarea acestui capitol, studentul va fi capabil să:
O6-1: să explice de ce este utilă eșantionarea unei populații și să descrie cele cinci metode principale de eșantionare
O6-2: să definească distribuțiile de sondaj și eroarea de sondaj
O6-3: să explice distribuția de sondaj a mediilor eșantioanelor
O6-4: să explice distribuția de sondaj a proporțiilor eșantioanelor
O6-5: să definească teorema limită centrală și să o aplice în calculul probabilităților
O6-6: să înțeleagă conceptele de bază ale eșantionării cu și fără revenire

6.1. AVANTAJELE EȘANTIONĂRII

Capitolele anterioare s-au focalizat cu precădere asupra tehnicilor de analiză descriptivă a datelor statistice (distribuții ale frecvențelor, indicatorii tendinței centrale și dispersiei), fiind studiate fenomene care au avut loc anterior. Ulterior au fost introduse primele elemente de statistică inferențială prin trecerea în revistă a noțiunilor de bază ale teoriei probabilităților, completate de prezentarea celor mai des întâlnite distribuții probabilistice discrete (binomială, hipergeometrică și Poisson) și continue (uniformă și normală). Principalul scop urmărit de statistica inferențială este reprezentat de obținerea unor informații pertinente despre anumite caracteristici ale unei populații pe baza datelor unui eșantion reprezentativ extras din aceasta. Distribuțiile probabilistice prezintă toate rezultatele posibile ale unui experiment împreună cu probabilitățile

Scanned with OKEN Scanner

asociate fiecăruia dintre acestea, fiind utilizate în evaluarea șanselor ca un anumit eveniment de interes să se producă în viitor.

În acest capitol, vom prezenta principalele noțiuni referitoare la procesul de eșantionare prin care se realizează selecția unităților statistice dintr-o populație în vederea utilizării acestor informații pentru a caracteriza întreaga colectivitate care prezintă interes pentru cercetare. Ulterior, vom construi distribuția mediilor eșantioanelor extrase pentru a înțelege modul în care mediile eșantioanelor tind să graviteze în jurul mediei populației din care provin. În partea finală, vom arăta faptul că, indiferent de tipul repartiției datelor din populația studiată, forma distribuției de sondaj a mediilor eșantioanelor de un anumit volum extrase din aceasta va tinde spre o distribuție probabilistică de tip Gauss-Laplace, pe măsura creșterii volumului eșantionului.

În cadrul studierii caracteristicilor unei populații, pentru a afla valoarea anumitor indicatori statistici (media, proporția, varianța, etc.) există o serie de considerente de natură practică datorită cărora preferăm să selectăm, observăm și măsurăm anumite porțiuni (eșantioane) ale colectivității generale, care să prezinte „în miniatură" trăsăturile esențiale ale acesteia.

Principalele avantaje ale procesului de eșantionare sunt:

- *perioada de timp mult mai redusă*: contactarea și măsurarea caracteristicii statistice din programul cercetării la nivelul întregii populații ar putea implica o perioadă imensă de timp (de ex., dacă un potențial candidat la președinția țării ar dori să determine șansele de a fi ales, acesta poate apela la o companie specializată în domeniul sondajelor de opinie, care va intervieva un eșantion reprezentativ de persoane, operațiune care ar dura câteva zile; în situația în care ar dori ca aceeași firmă de specialitate să investigheze toată populația țării cu drept de vot, ar fi nevoie de câteva luni pentru a finaliza cercetarea);
- *costuri mai mici*: cheltuielile implicate de studierea tuturor unităților statistice care compun populația generală ar putea fi imposibil de achitat (de ex., pentru a testa un produs nou, utilizarea unui eșantion reprezentativ de 1000 de persoane la care să fie trimise mostrele și colectate apoi rezultatele reprezintă un obiectiv tangibil, comparativ cu testarea produsului pe întreaga populație a țării);
- *imposibilitatea fizică de a verifica toate unitățile statistice* dintr-o anumită populație (anumite populații pot fi considerate ca fiind „infinite"; este imposibilă de ex. verificarea întregii cantități de apă dintr-un anumit lac pentru a determina nivelul bacteriilor, soluția fiind în schimb utilizarea mai multor eșantioane din diferite locații ale acestuia);
- *natura distructivă a unor teste* (de ex. dacă într-o anumită casă de vinuri se dorește o verificare a calității produselor, iar persoanele care efectuează testarea ar utiliza în analiză toată populația de sticle, nu ar mai rămâne niciun produs disponibil în vederea comercializării);

Scanned with OKEN Scanner

- *rezultatele obținute din analiza unui eșantion reprezentativ sunt adecvate* (chiar dacă de ex. resursele financiare, de timp și umane ar fi disponibile nelimitat, în vederea determinării indicelui lunar al prețurilor de consum pentru produsele alimentare, analiza tuturor magazinelor de specialitate din România nu ar afecta semnificativ valoarea indicelui, deoarece prețurile laptelui, pâinii și altor bunuri de consum alimentare nu înregistrează variații însemnate între marile lanțuri de magazine).

6.2. PRINCIPALELE METODE DE EȘANTIONARE

Așa cum am văzut anterior, rolul statisticii descriptive este centralizarea și prezentarea într-o formă convenabilă a ceea ce deja cunoaștem, iar cel al statisticii inferențiale este „să învățăm elemente noi, necunoscute încă, pornind de la ceea ce cunoaștem". Dacă teoria probabilităților reprezintă fundația care stă la baza întregi teorii statistice, teoria sondajelor poate fi privită ca fiind elementul structural esențial necesar pentru definitivarea întregii forme a construcției cunoașterii statistice. Astfel, teoria sondajelor deține un rol important în precizarea ipotezelor de lucru pe care se bazează analiza statistică inferențială ulterioară.

1. Eșantionarea aleatorie simplă

Sondajul întâmplător simplu se caracterizează prin următoarele trăsături specifice:

- fiecare unitate a populației are asigurată aceeași probabilitate de a face parte din eșantion;
- fiecare extragere a unei unități statistice din cadrul populației este independentă de oricare alta;
- volumul populației generale rămâne același pe parcursul tuturor extragerilor de constituire a eșantionului.

Principiile de constituire a eșantionului în cazul unui sondajului aleatoriu simplu se bazează pe schema probabilistică clasică a urnei lui Bernoulli cu bila revenită sau cea a bilei nerevenite (schema hipergeometrică). În funcție de schema probabilistică ce se utilizează pentru constituirea eșantionului, sondajul aleatoriu simplu poate fi repetat sau nerepetat.

În cazul sondajului aleatoriu simplu *repetat*, fiecare element component al populației (obiect, persoană, etc.) de volum N are aceeași șansă de a fi selectat și introdus în eșantionul de volum n, respectiv $p = 1 / N$.

Sondajul aleatoriu simplu *nerepetat* are loc atunci când unitatea statistică extrasă nu mai revine în cadrul populației N din care a fost extrasă. Aceasta înseamnă că, pe măsura constituirii eșantionului, volumul populației se diminuează cu unitățile selectate. Astfel, după prima extragere, populația va fi compusă din $N - 1$ elemente, după a doua extragere, din $N - 2$ și așa mai

Scanned with OKEN Scanner

departe, încât după extracția n vor rămâne în cadrul colectivității generale $N - n$ unități. Probabilitatea de selecție crește ușor pe măsura constituirii eșantionului, astfel, pentru prima unitate este $1 / N$, pentru a doua $1 / (N - 1)$, iar pentru extracția n va fi $1 / (N - n + 1)$.

În practică, o distribuție hipergeometrică poate fi de obicei aproximată printr-o distribuție binomială, întrucât în situația în care volumul eșantionului nu depășește 5% din cel al populației, există o diferență redusă între eșantionarea cu și fără revenire.

De exemplu, dintr-o populație de 2000 de studenți din anul I ai unei universități, dorim să extragem un eșantion reprezentativ de 50 de persoane. Putem nota numele fiecărui student, din cei 2000 care compun populația de referință, pe câte o bucată de hârtie și introduce aceste hârtii într-o cutie, unde, după ce le amestecăm bine, putem extrage cei 50 de studenți care vor constitui eșantionul de lucru, fără să returnăm fiecare bucată de hârtie în cutie. Observăm faptul că probabilitatea selectării fiecărui element nou în eșantion crește foarte puțin, datorită selecției cu revenire, dar volumul eșantionului fiind sub 5% din cel al populației ($50/2000 = 2{,}5\%$) diferențele nu influențează semnificativ rezultatele. Ca variantă alternativă, pentru selectarea eșantionului pot fi utilizate tabelele cu numere aleatoare, după realizarea în prealabil a unei liste cu cei 2000 de studenți. Se alege la întâmplare o căsuță din tabel pentru primul student extras și ulterior se selectează și restul de studenți din componența eșantionului, alegând în ordinea în care apar în tabelul cu numere aleatorii toate numerele cuprinse între 1 și 2000.

În acest fel, eșantionul va fi reprezentativ pentru populația de referință, putând fi astfel utilizat în analiza statistică ulterioară. În practică, este relativ ușor să extragem un eșantion dintr-o populație de referință, prin utilizarea programului de calcul tabelar Microsoft Excel (Modulul „Data Analysis" → „Sampling": la Input Range se introduce lista cu numerele de ordine ale elementelor populației generale, apoi se precizează metoda de eșantionare „Random" și volumul dorit al eșantionului „Number of Samples").

2. Eșantionarea aleatorie sistematică

Poate reprezenta o alternativă la eșantionarea aleatorie simplă în situația când dispunem de o listă completă a tuturor elementelor care compun o populație. În realizarea efectivă a selecției, se alege un punct de start (oricare element al listei) și prin adunarea unei constante la acesta se obține următorul element din eșantion, și așa mai departe, procedeul repetându-se până la completarea întregului volum al lotului de studiu. Observăm o anumită periodicitate în selectarea elementelor componente ale eșantionului, care în final, în anumite situații, poate conduce la erori de reprezentativitate a lotului extras. În cazul când lista care conține toate unitățile populației pe care o utilizăm prezintă și ea o ciclicitate asemănătoare, eșantioanele generate pot avea erori de deplasare (un *estimator este nedeplasat*, dacă valoarea așteptată a distribuției

Scanned with OKEN Scanner

sale de sondaj este egală cu valoarea reală, dar necunoscută a parametrului populației pe care acesta îl estimează, de exemplu, media mediilor tuturor eșantioanelor de un anumit volum n este egală cu media reală a populației din care acestea au fost extrase, reprezentând un estimator nedeplasat al acesteia; în schimb, media varianțelor acestor eșantioane este întotdeauna mai mică decât varianța populației generale, constituind un *estimator deplasat* al acesteia, în special în cazul eșantioanelor de volum redus). Eșantionarea sistematică poate fi o soluție adecvată în situația când eșantionarea aleatorie simplă nu poate fi aplicată.

De exemplu, un spital dorește să studieze nivelul de colesterol al pacienților care vin la consultații. Lista pacienților este organizată în funcție de programările din baza de date a spitalului. Vor fi urmați următorii pași:

- ✓ se stabilește dimensiunea eșantionului; presupunem faptul că lista conține $N = 1000$ de pacienți și se dorește un eșantion de volum $n = 100$. Prin urmare, se va selecta fiecare al 10-lea pacient ($1000/100 = 10$);
- ✓ se alege în mod aleatoriu un număr între 1 și 10 ca punct de start (să considerăm că a fost ales numărul 4);
- ✓ în continuare, va fi inclus în studiu fiecare al 10-lea pacient de pe listă (4, 14, 24, 34, 44, 54 etc.).

3. *Eșantionarea aleatorie stratificată*

În situația când o anumită populație poate fi împărțită în mai multe grupuri sau straturi în funcție de o variabilă statistică de interes ne aflăm în situația eșantionării stratificate, în care eșantionul de lucru se obține prin selectarea în mod aleatoriu a unui subeșantion din fiecare strat. Păstrarea proporției straturilor din populație în cadrul eșantionului conduce la o îmbunătățire a preciziei comparativ cu eșantionarea simplă aleatorie.

Eșantionarea stratificată poate fi în anumite situații mai eficientă comparativ cea simplă aleatorie, în special când anumite subpopulații sunt rare. De exemplu, în studiul schizofreniei este indicat să se împartă populația de referință în 2 părți: schizofrenici și clinic sănătoși, iar apoi să se selecteze aleatoriu un număr egal de persoane din fiecare grup. Dacă am fi selectat în mod aleatoriu persoanele din populația generală, exista riscul ca eșantionul să conțină extrem de puține persoane cu schizofrenie, încât studiul să fi fost lipsit de relevanță.

4. *Eșantionarea de tip cluster (ciorchine)*

Este utilizată deseori în scopul reducerii costurilor privind eșantionarea unei populații răspândite pe o arie geografică extinsă. În acest scop, populația studiată este împărțită în zone (grupuri sau clustere) utilizând de obicei limitele naturale ale regiunilor respective. Ulterior se selectează în mod aleatoriu un anumit număr de clustere din care va fi extrasă de asemenea în mod aleatoriu fiecare unitate care va intra în componența eșantionului.

Scanned with OKEN Scanner

5. Eșantionarea convenabilă

Unitățile componente ale viitorului eșantion sunt alese într-un mod convenabil de către cercetător și nu selectate în mod aleatoriu din populația de interes. Reprezintă o metodă mai accesibilă și ușor de aplicat, care nu garantează însă reprezentativitatea eșantionului obținut în acest fel. Deoarece permite obținerea mai rapidă a rezultatelor are rolul de a ajuta la conturarea unei prime imagini asupra fenomenului studiat. Estimările pot prezenta erori mari de deplasare, motiv pentru care rezultatele vor fi folosite doar ca studii pilot, care ajută la schițarea unor noi direcții de cercetare.

6.3. DISTRIBUȚIILE DE SONDAJ ȘI EROAREA DE SONDAJ

Conform *legii numerelor mari*, formulată pentru prima dată riguros de către matematicianul elvețian Jacob Bernoulli în lucrarea sa *„Ars Conjectandi" (Arta de a presupune)* publicată în anul 1713, dacă un experiment se repetă, în aceleași condiții, de un număr suficient de mare de ori, atunci frecvența relativă a unui eveniment prezintă o anumită stabilitate, variind în jurul probabilității de realizare a acestuia, bazată pe definiția clasică (Shafer Glenn, The significance of Jacob Bernoulli's Ars Conjectandi for the philosophy of probability today).

Aplicând această lege mediilor eșantioanelor - media putând fi privită ca frecvența de apariție a unui eveniment - se poate afirma faptul că pe măsură ce volumul eșantionului devine mai mare ($n \to \infty$), media acestuia tinde să se apropie de valoarea adevărată a mediei populației din care acesta provine ($\bar{x} \to \mu$).

Din punct de vedere tehnic, Legea numerelor mari poate fi aplicată oricărui *indicator* statistic calculat la nivelul unui eșantion sub forma unei *medii de valori independente*, cum ar fi de exemplu varianța eșantionului (media pătratelor abaterilor fiecărui nivel individual de la nivelul mediu al eșantionului – momentul de ordinul 2), asimetria (momentul de ordinul 3), boltirea (momentul de ordinul 4) sau coeficientul de corelație (media produsului abaterilor standard ale celor două variabile). Valorile minimă și maximă ale unui eșantion însă nu pot fi cuantificate sub forma unor mărimi medii și din acest motiv nu intră sub incidența legii numerelor mari.

În situația concretă a unui studiu trebuie însă să cunoaștem anumite elemente privitoare la comportamentul mediei eșantionului de lucru când aceasta este calculată dintr-un set de date de multe ori modest.

Distribuția de sondaj reprezintă mulțimea tuturor valorilor pe care le poate lua un indicator statistic calculat pentru toate eșantioanele de volum n, extrase în mod aleatoriu din cadrul aceleiași populații de volum N.

Construcția unei distribuții de sondaj se realizează extrem de dificil dacă populația de referință are un volum considerabil și reprezintă o sarcină imposibilă

Scanned with OKEN Scanner

dacă populația este infinită. În aceste situații, distribuțiile de sondaj vor putea fi aproximate prin extragerea unui număr mare de eșantioane având același volum n.

În cazul distribuțiilor de sondaj, suntem de obicei interesați să cunoaștem media, varianța și forma funcțională a acestora. Printre cele mai des întâlnite distribuții de sondaj, în practică întâlnim: distribuția mediilor de sondaj, distribuția proporțiilor de sondaj, distribuția diferențelor dintre 2 medii de sondaj și distribuția diferențelor dintre 2 proporții de sondaj.

Eșantioanele sunt utilizate pentru a estima caracteristicile populației supuse studiului, de exemplu media unui eșantion ($\bar{x}$) este folosită în vederea estimării mediei populației (μ) din care acesta provine. Având în vedere faptul că eșantionul reprezintă doar o mică parte a populației generale, este foarte puțin probabil ca media calculată la nivel de eșantion să fie egală cu cea a întregii populații. În mod similar, deviația standard calculată pentru datele eșantion (s) nu va fi perfect egală cu deviația standard a întregii populații (σ). Din acest considerent, ne așteptăm să existe o *diferență între statistica calculată la nivelul eșantionului și parametrul corespunzător care reprezintă întreaga colectivitate* sau populație de referință. Această diferență poartă numele de *eroare de sondaj*.

6.4. DISTRIBUȚIA DE SONDAJ A MEDIILOR EȘANTIOANELOR

Presupunem că avem o „populație fictivă" de $N = 120$ de persoane, pentru care dorim să analizăm nivelul de inteligență, cuantificat prin scorurile IQ. Variabila statistică IQ urmează o distribuție aproximativ normală (figura 6.1) având o medie $\mu = 100{,}39$ și o deviație standard de $\sigma = 15{,}27$. Nivelurile variabilei studiate sunt cuprinse între o valoare minimă de 67 și o valoare maximă de 132. Datele privind scorurile IQ ale celor 120 de persoane, care compun populația de referință, sunt ordonate crescător în tabelul 6.1.

Tabelul 6.1. Populația de 120 de persoane analizată prin prisma scorurilor IQ

67	76	85	90	95	99	102	104	107	111	116	122
68	78	85	91	95	99	102	104	108	112	117	123
69	78	86	91	95	100	102	104	109	112	117	124
71	79	86	92	96	100	103	104	109	113	117	124
72	80	87	93	96	100	103	105	109	113	118	125
73	81	88	93	96	100	103	105	109	114	119	127
73	82	88	93	97	100	103	106	110	114	120	129
75	83	89	93	98	101	103	106	110	115	121	130
76	84	89	94	98	101	103	106	111	115	121	131
76	84	90	95	98	101	103	107	111	115	121	132

Scanned with OKEN Scanner

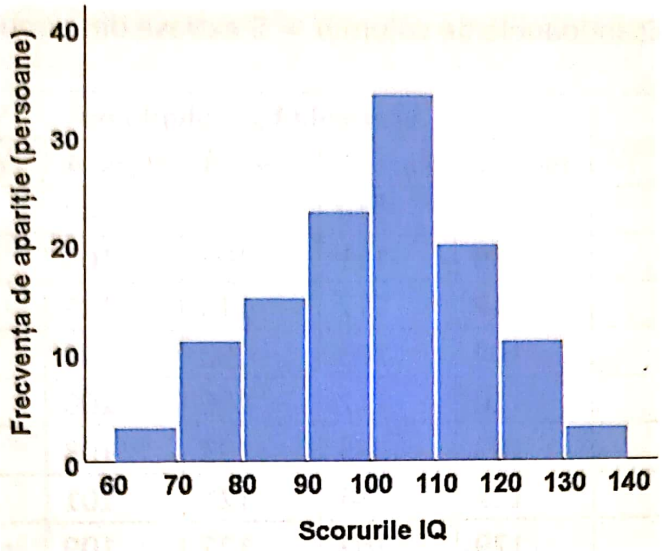


Figura 6.1. Histograma scorurilor IQ din populația de 120 de persoane

Distribuția de sondaj a mediilor eșantioanelor reprezintă o distribuție probabilistică a tuturor mediilor eșantioanelor de volum n extrase în mod aleatoriu dintr-o populație de volum N.

Din cadrul acestei populații de volum $N = 120$ de persoane analizate prin prisma scorurilor IQ putem extrage în mod aleatoriu un număr de:

$$C_{120}^{5} = \frac{120!}{5!\,(120-5)!} = 190.578.024$$

eșantioane posibile de volum $n = 5$ persoane.

Reprezentând grafic distribuția celor 190578024 valori medii calculate pentru toate eșantioane de volum $n = 5$ extrase aleatoriu din populația generală de 120 persoane, vom obține distribuția de sondaj a mediilor acestora.

Deoarece construirea în acest mod a distribuției de sondaj a mediilor eșantioanelor ar presupune un efort enorm, fiind aproape imposibil de realizat în practică, ne vom rezuma la utilizarea unui număr mai redus, de doar 10000 de eșantioane, având fiecare un volum $n = 5$ persoane, generate în mod aleatoriu prin utilizarea programului Microsoft Excel, de exemplu (tabelul 6.2).

Ulterior vom calcula media aritmetică a scorurilor IQ pentru fiecare dintre aceste 10000 de eșantioane de volum $n = 5$ (tabelul 6.2) și vom construi histograma tuturor acestor valori medii, care va contura o imagine aproximativă a distribuției de sondaj a tuturor celor 190578024 medii (figura 6.2).

Scanned with OKEN Scanner

Tabelul 6.2. Eșantioanele de volum $n = 5$ extrase din populația $N = 120$

Eșantioane	Scorurile IQ individuale					Media IQ
	Pers. 1	Pers. 2	Pers. 3	Pers. 4	Pers. 5	
Eșantion 1	93	119	100	87	118	103,4
Eșantion 2	89	111	102	103	98	100,6
Eșantion 3	117	117	119	115	95	112,6
Eșantion 4	104	103	102	90	87	97,2
Eșantion 5	110	112	106	100	108	107,2
Eșantion 6	103	85	93	108	92	96,2
Eșantion 7	117	84	121	101	83	101,2
Eșantion 8	129	102	121	109	115	115,2
Eșantion 9	92	125	110	121	100	109,6
Eșantion 10	95	110	117	91	93	101,2
...	...	...	...	...	...	...
Eșantion 9998	68	103	84	104	105	92,8
Eșantion 9999	103	83	87	117	129	103,8
Eșantion 10000	112	83	89	96	91	94,2

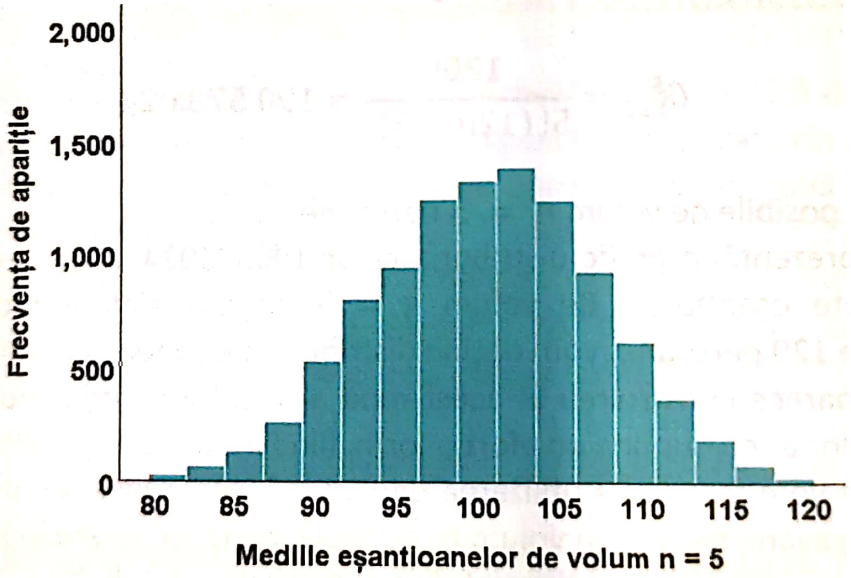


Figura 6.2. Distribuția de sondaj a mediilor eșantioanelor de volum $n = 5$

Observăm faptul că distribuția mediilor celor 10000 de eșantioane de volum $n = 5$ este aproximativ normală având o medie de 100,44, valoare care este aproape identică cu media populației generale din care acestea provin (100,39). Mediile fiecărui eșantion din cele 10000 extrase sunt cuprinse între o valoare minimă de 77,2 și o valoare maximă de 123,0. Observăm faptul că

Scanned with OKEN Scanner

experimentul nostru cu eșantioanele de volum $n = 5$ nu este foarte precis, mediile eșantioanelor având o variabilitate destul de mare în jurul mediei lor.

În situația în care utilizăm un eșantion mai consistent, de volum *n = 30*, media mediilor IQ ale celor 10000 de eșantioane va fi de 100,37 din nou extrem de apropiată de media populației generale (100,39) din care loturile au fost extrase (figura 6.3).

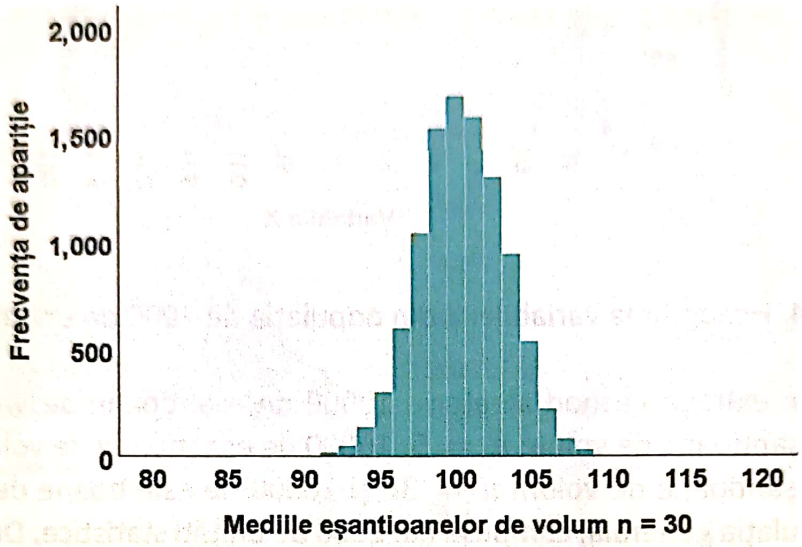


Figura 6.3. Distribuția de sondaj a mediilor eșantioanelor de volum $n = 30$

În același timp, remarcăm faptul că distribuția mediilor eșantioanelor este și în acest caz aproximativ normală, prezentând un nivel mult mai redus de împrăștiere a valorilor medii ale eșantioanelor comparativ cu cazul anterior și implicit o tendință de concentrare mai mare în jurul valorii reale a mediei populației de referință (variabila având o plajă de variație cuprinsă între 88,9 la 111,3). Dacă analizăm și valorile deviației standard în cele două cazuri, observăm o reducere importantă (de aproximativ 2,5 ori), de la un nivel de 6,84 în situația distribuției de sondaj a mediilor eșantioanelor de volum $n = 5$ până la 2,77 în cazul utilizării eșantioanelor de volum $n = 30$. De asemenea, în ambele situații deviațiile standard ale celor două distribuții de sondaj sunt mult mai mici comparativ cu deviația standard existentă în populația generală $\sigma = 15{,}27$.

Considerăm un al doilea exemplu, cel al unei populații compuse din $N = 2000$ de unități statistice, care urmează în acest caz o distribuție uniformă, fiind repartizată liniar constant în intervalul [60, 120], având o medie de 89,88 și o deviație standard de 17,92.

În vederea evidențierii din punct de vedere grafic a formei distribuției, construim și în acest caz histograma variabilei studiate X din populația generală (figura 6.4).

Scanned with OKEN Scanner

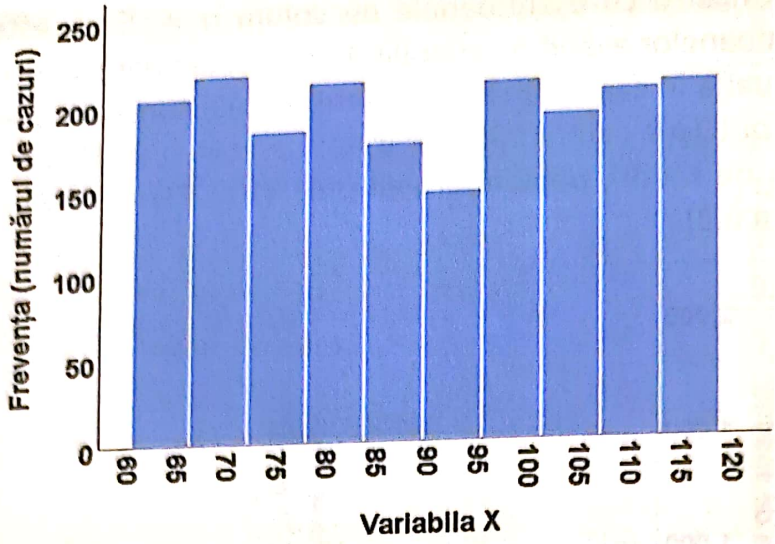


Figura 6.4. Histograma variabilei X din populația de 2000 de unități statistice

Vom extrage în mod aleatoriu 10000 de eșantioane de volum $n = 2$, 10000 de eșantioane de volum $n = 5$, 10000 de eșantioane de volum $n = 10$, 10000 de eșantioane de volum $n = 30$ și 10000 de eșantioane de volum $n = 50$ din populația generală, compusă din 2000 de unități statistice. Distribuțiile de sondaj ale mediilor celor 10000 de eșantioane, în funcție de volumul acestora, sunt prezentate în figura 6.5.

Și în al doilea exemplu observăm că distribuțiile de sondaj ale mediilor eșantioanelor sunt aproximativ normale în toate cele 5 cazuri, chiar dacă populația generală din care sunt extrase nu urmează o distribuție normală.

Media mediilor eșantioanelor ($\mu_{\bar{x}}$) este pentru fiecare situație în parte aproximativ egală cu cea a populației (μ). Dacă am putea dispune de toate eșantioanele de același volum extrase în mod aleatoriu din cadrul populației generale, atunci am avea relația:

$$\mu_{\bar{x}} = \mu \tag{6.1}$$

În cele cinci cazuri evidențiate anterior avem:

- $n = 2$ $\quad \mu_{\bar{x}} = 89{,}88 \approx \mu = 89{,}88;$
- $n = 5$ $\quad \mu_{\bar{x}} = 89{,}84 \approx \mu = 89{,}88;$
- $n = 10$ $\quad \mu_{\bar{x}} = 89{,}94 \approx \mu = 89{,}88;$
- $n = 30$ $\quad \mu_{\bar{x}} = 89{,}91 \approx \mu = 89{,}88;$
- $n = 50$ $\quad \mu_{\bar{x}} = 89{,}91 \approx \mu = 89{,}88.$

În schimb, deviația standard a mediilor eșantioanelor pentru fiecare caz în parte este în mod constant mai mică decât deviația standard a populației inițiale. De asemenea, observăm și în acest exemplu un nivel mult mai redus de

Scanned with OKEN Scanner

împrăștiere a valorilor medii ale eșantioanelor de la media mediilor acestora pe măsura creșterii volumului eșantionului de lucru.

- $n = 2$ $\quad \sigma_{\bar{x}} = 12{,}58 < \sigma = 17{,}92;$
- $n = 5$ $\quad \sigma_{\bar{x}} = 8{,}07 < \sigma = 17{,}92;$
- $n = 10$ $\quad \sigma_{\bar{x}} = 5{,}69 < \sigma = 17{,}92;$
- $n = 30$ $\quad \sigma_{\bar{x}} = 3{,}27 < \sigma = 17{,}92;$
- $n = 50$ $\quad \sigma_{\bar{x}} = 2{,}52 < \sigma = 17{,}92.$

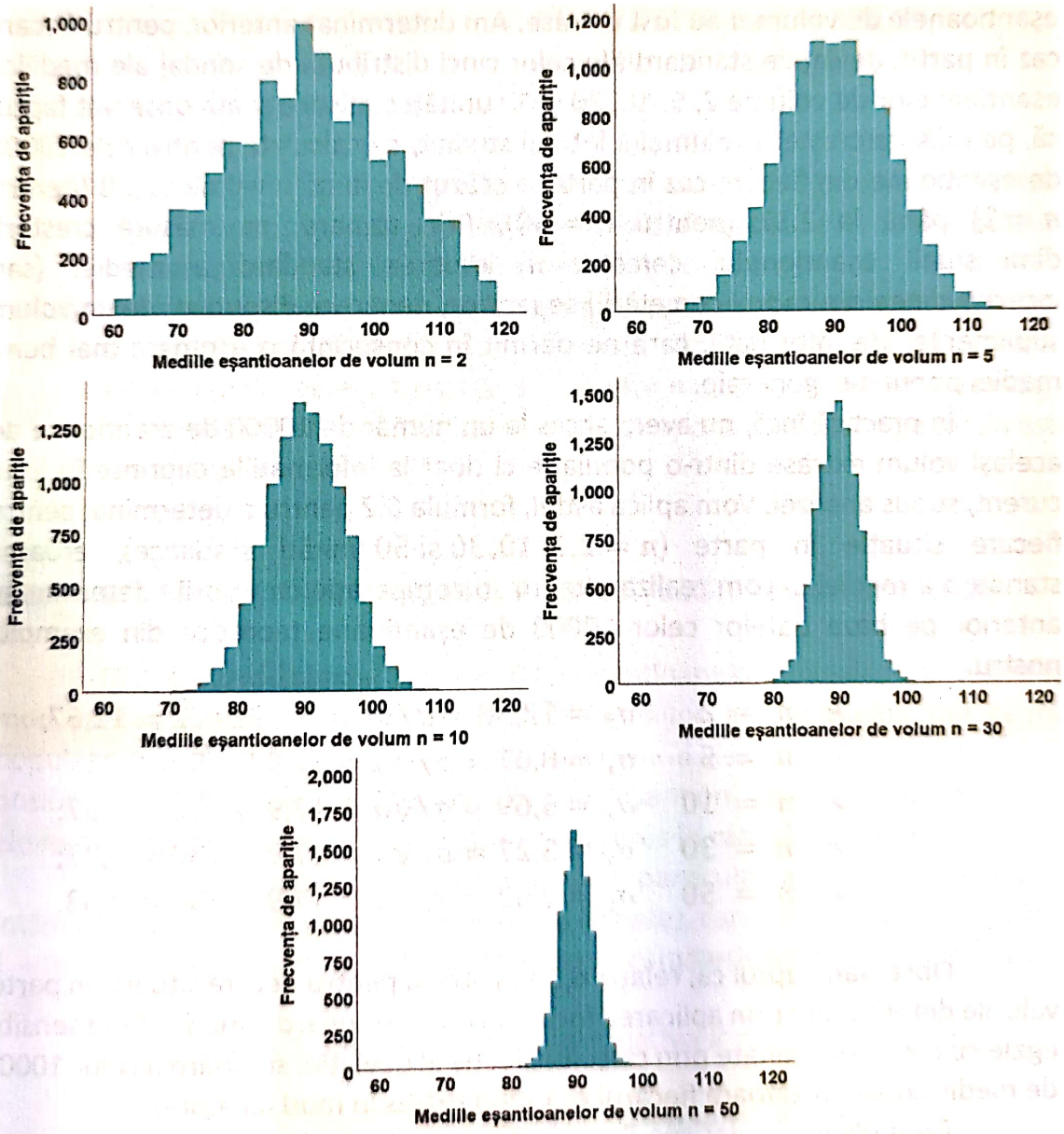

Frecvența de apariție
Mediile eșantioanelor de volum n = 2
Mediile eșantioanelor de volum n = 5
Mediile eșantioanelor de volum n = 10
Mediile eșantioanelor de volum n = 30
Mediile eșantioanelor de volum n = 50

Figura 6.5. Distribuțiile de sondaj ale mediilor eșantioanelor de volume 2, 5, 10, 30 și 50 de unități

Scanned with OKEN Scanner

Se poate demonstra faptul că, dacă într-o anumită populație avem o abatere standard σ, atunci deviația standard a distribuției de sondaj a tuturor mediilor eșantioanelor de volum n extrase în mod aleatoriu din aceasta $\sigma_{\bar{x}}$ va fi:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{6.2}$$

În statistică, $\sigma_{\bar{x}}$ mai poartă numele de *eroarea standard a mediei* și este de $\sqrt{n}$ ori mai mică decât deviația standard a populației generale din care eșantioanele de volum n au fost extrase. Am determinat anterior, pentru fiecare caz în parte, deviațiile standard ale celor cinci distribuții de sondaj ale mediilor eșantioanelor de volume 2, 5, 10, 30 și 50 unități statistice și am observat faptul că, pe măsura creșterii volumului lotului studiat, $\sigma_{\bar{x}}$ calculată pentru cele 10000 de eșantioane din fiecare caz în parte, a scăzut de la un nivel de 12,58 (pentru $n = 2$) până la 2,52 (pentru $n = 50$). Prin urmare, pe măsura creșterii dimensiunii eșantionului cercetat n, eroarea standard a mediei (sau incertitudinea din cadrul cercetării) se reduce, deoarece dispunem de un volum suplimentar de informații, care ne permit în consecință o estimare mai bună mediei populației generale.

În practică însă, nu avem acces la un număr de 10000 de eșantioane de același volum extrase dintr-o populație ci doar la informațiile cuprinse în lotul curent, supus analizei. Vom aplica astfel, formula 6.2 pentru a determina, pentru fiecare situație în parte ($n = 2, 5, 10, 30$ și 50 unități statistice), eroarea standard a mediei și vom realiza ulterior o comparație cu valorile determinate anterior pe baza datelor celor 10000 de eșantioane teoretice din exemplul nostru.

- $n = 2$ $\quad \sigma_{\bar{x}} = 12{,}58 \approx \sigma/\sqrt{n} = 17{,}92/\sqrt{2} = 12{,}67;$
- $n = 5$ $\quad \sigma_{\bar{x}} = 8{,}07 \approx \sigma/\sqrt{n} = 17{,}92/\sqrt{5} = 8{,}01;$
- $n = 10$ $\quad \sigma_{\bar{x}} = 5{,}69 \approx \sigma/\sqrt{n} = 17{,}92/\sqrt{10} = 5{,}67;$
- $n = 30$ $\quad \sigma_{\bar{x}} = 3{,}27 \approx \sigma/\sqrt{n} = 17{,}92/\sqrt{30} = 3{,}27;$
- $n = 50$ $\quad \sigma_{\bar{x}} = 2{,}52 \approx \sigma/\sqrt{n} = 17{,}92/\sqrt{50} = 2{,}53.$

Observăm faptul că, relația 6.2 se verifică pentru fiecare situație în parte, valorile determinate prin aplicarea formulei erorii standard a mediei fiind sensibil egale cu cele determinate prin calculul efectiv al deviației standard a celor 10000 de medii, corespunzătoare fiecărui eșantion extras în mod aleatoriu.

Exemplele de mai sus ilustrează o serie de relații importante care se conturează între distribuția populației generale și distribuția de sondaj a mediilor eșantioanelor de diferite volume extrase din aceasta:

a) *media mediilor eșantioanelor este aproximativ egală cu media populației generale* (în cazul în care am putea utiliza în calcul mediile tuturor

Scanned with OKEN Scanner

eșantioanelor de același volum extrase aleatoriu din cadrul populației de referință cele 2 valori ar fi perfect egale). Acest aspect atestă faptul că media mediilor eșantioanelor este un *estimator nedeplasat* al parametrului pe care îl estimează – media populației generale;

b) *media variațelor eșantioanelor în fiecare caz este mai mică comparativ cu varianța populației generale*, fapt care ne arată că acesta este un *estimator deplasat* al parametrului pe care îl estimează – varianța populației generale;

- $n = 2$ $\quad \bar{\sigma}^2 = 163{,}83 < \sigma^2 = 320{,}99;$
- $n = 5$ $\quad \bar{\sigma}^2 = 255{,}78 < \sigma^2 = 320{,}99;$
- $n = 10$ $\quad \bar{\sigma}^2 = 286{,}65 < \sigma^2 = 320{,}99;$
- $n = 30$ $\quad \bar{\sigma}^2 = 309{,}79 < \sigma^2 = 320{,}99;$
- $n = 50$ $\quad \bar{\sigma}^2 = 314{,}82 < \sigma^2 = 320{,}99.$

c) *varianța distribuției de sondaj a mediilor eșantioanelor este mai redusă decât cea existentă în populația generală*, fiind din ce în ce mai scăzută pe măsura creșterii volumului eșantionului de lucru;

d) *distribuția de sondaj a mediilor eșantioanelor poate fi aproximată printr-o distribuție probabilistică normală* în situația când volumul eșantioanelor este mare ($n \geq 30$). Pentru $n \leq 30$ distribuția de sondaj a mediilor poate fi considerată normală doar daca populația generală este normal distribuită.

6.5. DISTRIBUȚIA DE SONDAJ A PROPORȚIILOR EȘANTIOANELOR

Distribuția de sondaj a proporțiilor eșantioanelor se realizează în același mod cu cea prezentată anterior în cazul mediilor eșantioanelor. Din cadrul populației generale de volum N, vom extrage în mod aleatoriu toate eșantioanele posibile de volum n, iar pentru fiecare eșantion vom calcula proporția elementelor care prezintă caracteristica de interes în cadrul cercetării.

Proporția poate fi asimilată unui caz particular al mediei aritmetice, întâlnit în situația unei variabile alternative (binare), care poate lua doar 2 valori: 1 – în cazul variantelor variabilei care posedă caracteristica și 0 în rest (tabelul 6.3).

Media aritmetică a unei variabile alternative $\bar{x} = p$ (formula 6.3).

$$\bar{x} = \frac{\sum_{i=1}^{2} x_i \cdot n_i}{\sum_{i=1}^{2} n_i} = \frac{\sum_{i=1}^{2} x_i \cdot f_i}{\sum_{i=1}^{2} n_i} = \frac{1 \cdot p + 0 \cdot q}{p + q} = \frac{p}{1} = p \qquad (6.3)$$

Scanned with OKEN Scanner

Tabelul 6.3. Distribuția unei variabile alternative X

Variabila X	Frecvența absolută (n_i)	Frecvența relativă (f_i)
$x_1 = 1$	n_1	$n_1/(n_1+n_2) = p$
$x_2 = 0$	n_2	$n_2/(n_1+n_2) = q$
Total	$\boldsymbol{n_1+n_2}$	$\boldsymbol{p+q=1}$

Varianța unei variabile alternative $\sigma^2 = p \cdot (1-p)$ se determină astfel:

$$\sigma^2 = \frac{\sum_{i=1}^{2}(x_i - \bar{x})^2 \cdot f_i}{\sum_{i=1}^{2} f_i} = \frac{(1-p)^2 \cdot p + (0-p)^2 \cdot q}{p+q}$$

$$\sigma^2 = \frac{q^2 \cdot p + p^2 \cdot q}{p+q} = \frac{p \cdot q \cdot (p+q)}{p+q} = p \cdot q = p \cdot (1-p) \quad (6.4)$$

Considerăm o populație fictivă de $N = 720$ de persoane pentru care dorim să analizăm proporția adulților cu vârste între 45 și 50 de ani care suferă de prezbiopie. Prezbiopia reprezintă o afecțiune oculară comună, care apare odată cu înaintarea în vârstă, afectând capacitatea ochiului de a focaliza obiectele apropiate (tabelul 6.4).

Tabelul 6.4. Populația de 720 de persoane analizată în funcție de prezența prezbiopiei (1 – DA / 0 - NU)

1	1	1	1	1	1	1	1	1	1	...	1
1	0	1	0	1	0	1	0	1	0	...	0
1	1	1	1	1	1	1	1	1	1	...	1
0	1	0	1	0	1	0	1	0	0	...	1
1	1	1	1	1	1	1	1	0	1	...	0
1	0	1	0	1	0	1	0	1	0	...	0
1	1	1	1	1	1	1	1	1	1	...	0
0	1	0	1	0	1	0	1	0	1	...	0
1	1	1	1	1	1	1	1	1	0	...	0
1	0	1	0	1	0	1	0	1	0	...	0

Scanned with OKEN Scanner

Variabila statistică analizată este alternativă, fiecare individ putând fi încadrat, în funcție de prezența afecțiunii doar într-o singură categorie din două posibile (1 - cei care suferă de prezbiopie, 0 – persoanele sănătoase). La nivelul populației generale, proporția persoanelor care au afecțiunea (1) a fost de 0,55 sau 55% din total.

Din cadrul populației generale „fictive" de 720 de persoane putem extrage în mod aleatoriu un număr de:

$$C_{720}^{12} = \frac{720!}{12!\,(720-12)!} = 3.6951E+25 = 3.6951 \times 10^{25}$$

eșantioane posibile de volum $n = 12$ persoane.

Reprezentând grafic distribuția celor 3.6951×10^{25} valori ale proporțiilor, calculate pentru toate eșantioane de volum $n = 12$, extrase aleatoriu din populația generală de 720 persoane, vom obține distribuția de sondaj a proporțiilor acestora. Deoarece construirea în acest mod a distribuției de sondaj a proporțiilor eșantioanelor este imposibil de realizat în practică, ne vom rezuma la utilizarea unui număr mai redus, de doar 10000 de eșantioane, având fiecare un volum $n = 12$ persoane, generate în mod aleatoriu.

Ulterior vom calcula proporția persoanelor care prezintă prezbiopie pentru fiecare dintre aceste 10000 de eșantioane de volum $n = 12$ (tabelul 6.5) și vom construi histograma tuturor acestor valori ale proporțiilor, care va contura o imagine aproximativă a distribuției de sondaj a tuturor celor 10000 de valori ale proporțiilor (figura 6.6).

Tabelul 6.5. Eșantioanele de volum $n = 12$ extrase din populația $N = 720$

Eșantioane	**Persoane din eșantion care prezintă sau nu prezbiopie**								**Proporția *„media"***
	P1	**P2**	**P3**	**P4**	**P5**	**...**	**P11**	**P12**	
Eșantion 1	0	1	1	0	1	...	1	1	0,58
Eșantion 2	0	1	1	0	0	...	0	1	0,50
Eșantion 3	1	1	1	0	1	...	1	0	0,83
Eșantion 4	1	1	1	1	1	...	1	0	0,58
Eșantion 5	1	1	1	1	0	...	1	1	0,83
....	...	...	...	...	...	...	...	...	...
Eșantion 9999	1	1	1	1	1	...	0	1	0,75
Eșantion 10000	1	0	0	1	0	...	0	1	0,42

Scanned with OKEN Scanner

Observăm faptul că distribuția proporțiilor celor 10000 de eșantioane de volum n = 12 este aproximativ normală având o medie de 0,55, valoare care este aproape identică cu proporția populației generale din care acestea provin.

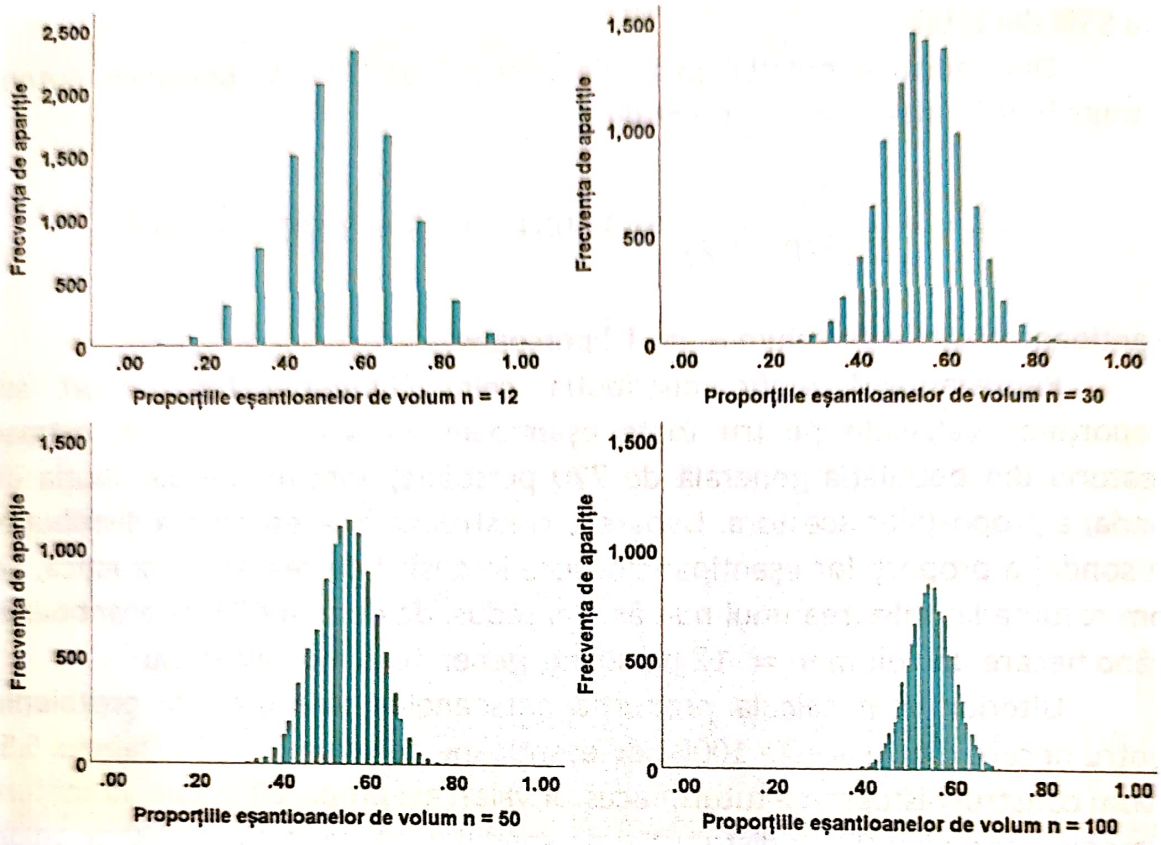


Figura 6.6. Distribuțiile de sondaj ale proporțiilor eșantioanelor de volume 12, 30, 50 și 100 de persoane cu vârstele între 45 și 50 de ani

Proporțiile determinate pentru fiecare eșantion din cele 10000 extrase sunt cuprinse între o valoare minimă de 0 (toate persoanele sunt sănătoase) și o valoare maximă de 1 (toate persoanele suferă de prezbiopie). Observăm faptul că experimentul nostru cu eșantioanele de volum n = 12 nu este foarte precis, proporțiile eșantioanelor având o variabilitate foarte mare în jurul valorii centrale (0,55).

Repetăm experimentul pentru eșantioane de volum 30, 50 și 100 (figura 6.6) și observăm faptul că media (valoarea așteptată) a proporțiilor eșantioanelor este în fiecare caz în parte foarte apropiată de proporția populației generale de 720 de persoane din care loturile au fost extrase (0,55).

În același timp, remarcăm faptul că distribuția proporțiilor eșantioanelor este și în acest cazuri aproximativ normală, prezentând un nivel din ce în ce mai redus de împrăștiere în jurul mediei acestora pe măsura creșterii volumului eșantionului, comparativ cu cazul anterior, și implicit o tendință de concentrare mai mare în jurul valorii reale a proporției populației de referință (figura 6.6).

Scanned with OKEN Scanner

Similar cu situația distribuției de sondaj a mediilor eșantioanelor și pentru distribuția de sondaj a proporțiilor eșantioanelor există o serie de relații importante care se conturează între distribuția populației generale și distribuția de sondaj a proporțiilor eșantioanelor de diferite volume extrase din aceasta:

a) media proporțiilor eșantioanelor este aproximativ egală cu proporția populației generale (în cazul în care am putea utiliza în calcul proporțiile tuturor eșantioanelor de același volum extrase în mod aleatoriu din cadrul populației de referință cele 2 valori ar fi perfect egale). Acest aspect atestă faptul că media proporțiilor tuturor eșantioanelor de același volum n extrase dintr-o populație de volum N este un *estimator nedeplasat* al parametrului pe care îl estimează – proporția din populația generală;
b) media variațelor eșantioanelor în fiecare caz este mai mică comparativ cu varianța populației generale, fapt care ne arată că acesta este un *estimator deplasat* al parametrului pe care îl estimează – varianța proporției populației generale;
c) varianța distribuției de sondaj a proporțiilor este mai redusă decât cea existentă în populația generală, fiind din ce în ce mai scăzută pe măsura creșterii volumului eșantionului de lucru;
d) distribuția de sondaj a proporțiilor eșantioanelor poate fi aproximată printr-o distribuție probabilistică normală în situația când volumul eșantioanelor este mare; condiția este îndeplinită când sunt satisfăcute ambele relații: $n \cdot p \geq 5$ și $n \cdot (1-p) \geq 5$.

6.6. TEOREMA LIMITĂ CENTRALĂ

A. Distribuția de sondaj a mediilor

Atunci când extragem eșantioane de un anumit volum n din cadrul oricărei populații de volum N, indiferent de tipul de distribuție al acesteia, distribuția de sondaj a tuturor mediilor eșantioanelor tinde către repartiția normală (Gauss-Laplace), aproximarea fiind din ce în ce mai bună pe măsura creșterii volumului acestora ($n \geq 30$).

- Dacă populația generală de volum N are o medie μ și o deviație standard σ, atunci distribuția de sondaj a mediilor tuturor eșantioanelor de volum n extrase aleatoriu din aceasta are de asemenea o medie μ și o eroare standard SEM (deviație standard a distribuției de sondaj a mediilor eșantioanelor $\sigma_{\bar{x}}$) , care se determină conform relației 6.2;
- Întotdeauna avem nivel mai redus al dispersiei (variației) în distribuția de sondaj a mediilor eșantioanelor comparativ cu cel înregistrat în populația de referință. Observăm faptul că pe măsura creșterii volumului eșantionului de studiu n, eroarea standard a mediei $\sigma_{\bar{x}}$ se reduce.

Scanned with OKEN Scanner

Eroarea standard a fost determinată anterior având la bază presupunerea faptului că eșantionarea s-a efectuat cu repetare sau eșantionul a fost extras dintr-o populație „infinită". În general, procesul de eșantionare se efectuează fără revenire și în cele mai multe situații din cadrul unei populații de volum finit.

Astfel, atunci când eșantionul este extras fără revenire dintr-o populație finită de volum N, distribuția de sondaj a mediilor tuturor eșantioanelor de volum n extrase aleatoriu din aceasta va avea media μ și eroarea standard $\sigma_{\bar{x}}$:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}} \tag{6.5}$$

Elementul care apare în acest caz suplimentar în formulă, respectiv $\sqrt{\frac{N-n}{N-1}}$, poartă denumirea de *factor de corecție* pentru o populație finită și poate fi ignorat atunci când volumul eșantionului este foarte mic în raport cu cel al populației din care provine (de obicei când $n/N \leq 0{,}05$).

B. *Distribuția de sondaj a proporțiilor*

În mod similar, în situația când extragem eșantioane de un anumit volum n din cadrul oricărei populații de volum N, indiferent de tipul de distribuție al acesteia, distribuția de sondaj a tuturor proporțiilor eșantioanelor tinde către repartiția normală, în situația când volumul eșantioanelor este suficient de mare ($n \cdot p \geq 5$ și $n \cdot (1-p) \geq 5$).

- Dacă populația generală de volum N are o proporție p (formula 6.3) și o deviație standard $\sigma = \sqrt{p \cdot (1-p)}$ (formula 6.4), atunci distribuția de sondaj a proporțiilor tuturor eșantioanelor de volum n extrase aleatoriu din aceasta are de asemenea o proporție p și o eroare standard a proporțiilor SEP (deviație standard a distribuției de sondaj a proporțiilor eșantioanelor $\sigma_{\bar{p}}$) egală cu:

$$\sigma_{\bar{p}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p \cdot (1-p)}{n}} \tag{6.6}$$

- Întotdeauna avem nivel mai redus al dispersiei în distribuția de sondaj a proporțiilor eșantioanelor comparativ cu cel înregistrat în populația de referință. Observăm faptul că pe măsura creșterii volumului eșantionului de studiu n, eroarea standard a proporției se reduce.

În mod similar cu distribuția de sondaj a mediilor și în acest caz, dacă analizăm o populație finită de volum N în calculul erorii standard a proporției se va aplica factorul de corecție pentru o populație finită, respectiv $\sqrt{\frac{N-n}{N-1}}$.

Scanned with OKEN Scanner

6.7. APLICAȚII PRACTICE ALE DISTRIBUȚIEI DE SONDAJ A MEDIILOR EȘANTIOANELOR

În scopul unei înțelegeri aprofundate a principalelor noțiuni ale statisticii inferențiale, care vor fi detaliate în capitolele viitoare, este necesară cunoașterea principalelor aplicații practice ale distribuțiilor de sondaj. Cea mai simplă dintre acestea, se referă la determinarea probabilității de a obține un eșantion de volum n cu o medie de o anumită magnitudine, extras dintr-o populație generală compusă din N unități statistice, pentru care media și deviația standard sunt cunoscute.

Având la dispoziție informații prealabile despre media μ, abaterea standard a unei populații σ și dimensiunea unui eșantion extras în mod aleatoriu din aceasta n, putem obține distribuția de sondaj a mediilor tuturor eșantioanelor de același volum extrase din populația generală și în același timp putem calcula și probabilitatea ca media aritmetică a eșantionului studiat să fie situată într-un anumit interval.

Așa cum am văzut anterior, *distribuția de sondaj a mediilor eșantioanelor tinde către repartiția normală* în cazul în care:

- eșantioanele sunt extrase din populații normal distribuite, situație în care volumul eșantionului nu joacă un rol important;
- atunci când forma distribuției populației generale nu este cunoscută, dimensiunea eșantionului joacă un rol important; în practică distribuția de sondaj a mediilor se va apropia de o distribuție normală, în situația extragerii unor eșantioane de cel puțin 30 de observații.

Din capitolul precedent cunoaștem faptul că *orice distribuție probabilistică normală poate fi convertită într-o distribuție normală standard* prin scăderea mediei din fiecare nivel individual al variabilei analizate și raportarea acestei diferențe la deviația standard a seriei:

$$z_i = \frac{x_i - \mu}{\sigma}$$

În practică însă suntem interesați însă de distribuția mediilor eșantioanelor $\bar{x}$ în jurul mediei populației generale μ, în locul distribuției variantelor individuale ale variabilei x_i în jurul mediei lor $\bar{x}$. În al doilea rând, în cadrul acestei distribuții a mediilor eșantioanelor vom utiliza în locul deviației standard a populației generale σ, deviația standard a distribuției de sondaj a mediilor tuturor eșantioanelor extrase aleatoriu din aceasta sau eroarea standard a mediei $\sigma_{\bar{x}}$:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Scanned with OKEN Scanner

Singura valoare care rămâne în aceeași formă în cadrul formulei noi este *media mediilor tuturor eșantioanelor* de volum n ($\mu_{\bar{x}}$) extrase în mod aleatoriu din populația de volum N, care reprezintă un *estimator nedeplasat al mediei populației generale* (μ), având aceeași valoare cu aceasta ($\mu_{\bar{x}} = \mu$).

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \tag{6.7}$$

Astfel, atunci când dorim să aflăm probabilitatea ca media eșantionului $\bar{x}$ extras dintr-o anumită populație, având media μ și deviația standard σ cunoscute, să fie cuprinsă într-un anumit interval, vom calcula valoarea $z_{\bar{x}}$ aferentă, iar apoi vom citi din tabelele distribuției normale valoarea probabilității corespunzătoare acesteia

Exemplu: Cunoaștem faptul că, într-o anumită populație valorile acidului uric pentru persoanele adulte de sex masculin, clinic sănătoase, sunt aproximativ normal distribuite cu o medie $\mu = 5{,}7$ mg și o deviație standard de $\sigma = 1$ mg și dorim să calculăm probabilitățile ca un eșantion de 30 de persoane de sex masculin extras în mod aleatoriu din cadrul acesteia să aibă media mai mare decât 6 mg, cuprinsă între 5 si 6 mg si mai mică decât 5 mg.

Avem următoarele date pentru nivelul acidului uric la bărbații adulți clinic sănătoși:

- ✓ Variabila X (valorile acidului uric) în populația de bărbați clinic sănătoși urmează o distribuție normală;
- ✓ Volumul eșantionului extras: $n = 30$ persoane de sex masculin;
- ✓ Media populației generale: $\mu = 5{,}7$ mg;
- ✓ Deviația standard a populației: $\sigma = 1$ mg.

Conform teoremei limită centrală, distribuția tuturor mediilor eșantioanelor de volum $n = 30$ persoane, extrase în mod aleatoriu din această populație, va fi aproximativ normală, având media μ și deviația standard $\sigma_{\bar{x}} = {}^{\sigma}/_{\sqrt{n}}$.

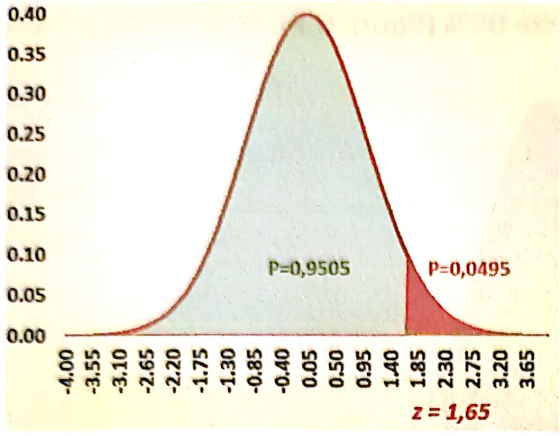
Pentru a determina probabilitatea ca eșantionul extras să aibă media acidului uric mai mare de 6 mg, calculăm valoarea corespunzătoare a lui z pentru media eșantionului nostru ($\bar{x}$) utilizând formula 6.7:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{6 - 5{,}7}{{}^{1}/_{\sqrt{30}}} \approx 1{,}65$$

Media eșantionului nostru se abate de la media mediilor tuturor eșantioanelor de volum $n = 30$ sau media întregii populații cu 1,65 deviații standard la dreapta (valoarea lui z exprimă eroarea de sondaj $\bar{x} - \mu$ în unități

Scanned with OKEN Scanner

standardizate). În continuare, vom calcula șansa obținerii unei valori z mai mari decât 1,65 (figura 6.7).



DISTRIBUTIA NORMALA STANDARD

z	Probabilitati cumulate						
	0.00	0.01	0.02	0.03	0.04	0.05	0.06
0.00	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239
0.10	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636
0.20	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026
0.30	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406
...	...	...	...	...	...	...	...
1.40	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279
1.50	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406
1.60	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515
1.70	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608

Figura 6.7. Distribuția normală standard $P(z) > 1{,}65$

Folosind tabelele distribuției normale standard (figura 6.7), avem:

$$P(z > 1{,}65) = 1 - P(z \leq 1{,}65) = 1 - 0{,}9505 = 0{,}0495$$

În concluzie, probabilitatea de a extrage în mod aleatoriu un eșantion de bărbați clinic sănătoși, având media acidului uric peste 6 mg, din cadrul unei populații cu media de 5,7 mg și deviația standard de 1 mg este extrem de redusă, fiind situată sub nivelul de 5%.

Pentru a determina probabilitatea ca eșantionul extras să aibă media acidului uric cuprinsă între 5 și 6 mg, calculăm valorile corespunzătoare ale scorurilor z în cele 2 situații:

$$z_1 = \frac{\bar{x}_1 - \mu}{\sigma/\sqrt{n}} = \frac{5 - 5{,}7}{1/\sqrt{30}} \approx -3{,}85$$

$$z_2 = \frac{\bar{x}_2 - \mu}{\sigma/\sqrt{n}} = \frac{6 - 5{,}7}{1/\sqrt{30}} \approx 1{,}65$$

Prin urmare:

$$P(5 \leq \bar{x} \leq 6) = P(-3{,}85 \leq z \leq 1{,}65) = P(z \leq 1{,}65) - P(z \leq -3{,}85) =$$

$$= 0{,}9505 - 0{,}0001 = 0{,}9504$$

Scanned with OKEN Scanner

În urma analizei, putem afirma faptul că, probabilitatea să extragem în mod aleatoriu un eșantion de bărbați clinic sănătoși, având media acidului uric cuprinsă între 5 și 6 mg, din cadrul unei populații cu media de 5,7 mg și deviația standard de 1 mg este foarte ridicată, peste 95% (figura 6.8).

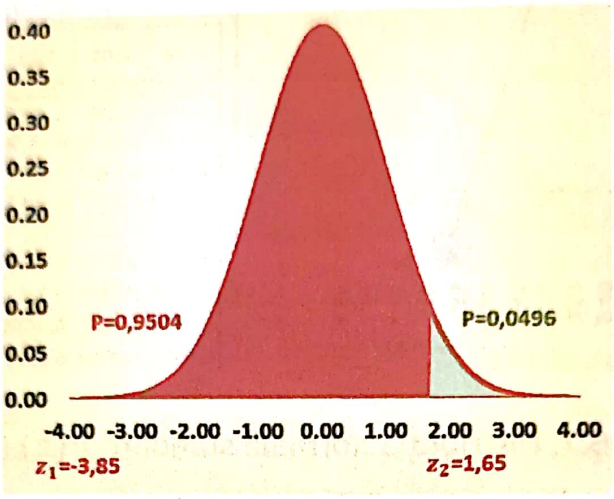


Figura 6.8. Distribuția normală standard $P(-3{,}85 \leq z \leq 1{,}65)$

În final, pentru a calcula șansa ca eșantionul extras să aibă media acidului uric mai mică de 5 mg, calculăm valoarea corespunzătoare a lui z pentru media eșantionului nostru ($\bar{x}$) utilizând formula 6.7:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{5 - 5{,}7}{1/\sqrt{30}} \approx -3{,}85$$

$$P(\bar{x} < 5) = P(z < -3{,}85) \approx 0{,}0001$$

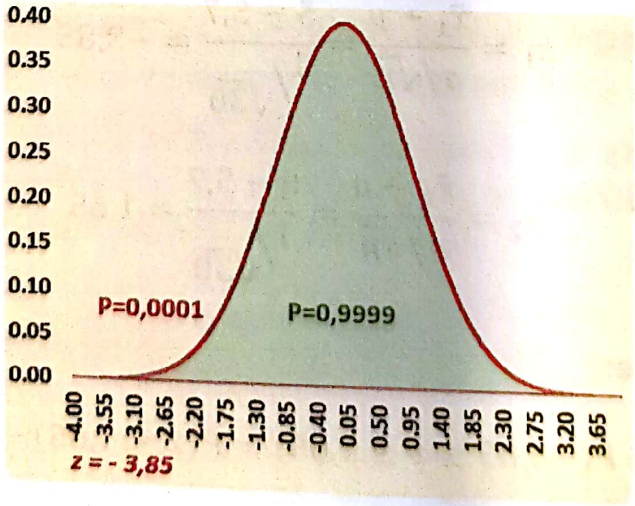


Figura 6.9. Distribuția normală standard $P(z) < -3{,}85$

Scanned with OKEN Scanner

Așa cum este evidențiat și în figura 6.9, probabilitatea să extragem în mod aleatoriu un eșantion de bărbați clinic sănătoși, având media acidului uric sub nivelul de 5 mg, din cadrul unei populații cu media de 5,7 mg și deviația standard de 1 mg este extrem de redusă, fiind situată cu mult sub 1% (figura 6.9).

Scanned with OKEN Scanner

Capitolul. 7

ESTIMAREA ȘI INTERVALELE DE ÎNCREDERE

SUBIECTE:

7.1. Estimarea punctuală și prin interval de încredere a parametrilor unei populații
7.2. Intervalul de încredere pentru estimarea mediei unei populații
7.3. Intervalul de încredere pentru estimarea diferenței dintre mediile a două populații
7.4. Intervalul de încredere pentru estimarea proporției unei populații
7.5. Intervalul de încredere pentru estimarea diferenței dintre proporțiile a două populații

La finalizarea acestui capitol, studentul va fi capabil să:
O7-1: să calculeze și să interpreteze o estimare punctuală și prin interval de încredere a mediei unei anumite populații
O7-2: să calculeze și să interpreteze o estimare punctuală și prin interval de încredere a proporției unei anumite populații
O7-3: să calculeze și să interpreteze o estimare prin interval de încredere a diferenței dintre mediile a două populații
O7-4: să calculeze și să interpreteze o estimare prin interval de încredere a diferenței dintre proporțiile a două populații

7.1. ESTIMAREA PUNCTUALĂ ȘI PRIN INTERVAL DE ÎNCREDERE A PARAMETRILOR UNEI POPULAȚII

În capitolul precedent am trecut în revistă principalele metode de eșantionare (simplă aleatoare – în care fiecare membru al populației generale are aceeași șansă de a fi selectat în eșantion, sistematică, stratificată, tip cluster și de convenienţă) și motivele care stau la baza acestui proces de selectare a unităților statistice dintr-o populație (timp îndelungat, costuri ridicate, rezultatele la nivel de eșantion sunt adecvate, natura distructivă a unor teste, imposibilitatea verificării tuturor observațiilor).

În cadrul exemplelor privind distribuțiile de sondaj ale mediilor eșantioanelor am considerat faptul că media, deviația standard și forma distribuției populației de referință sunt cunoscute. În cele mai multe situații reale care pot apare în practică însă asemenea informații nu sunt disponibile, principalul scop al cercetării statistice prin sondaj fiind estimarea acestor

Scanned with OKEN Scanner

parametri ai populației studiate în baza statisticilor calculate la nivel de eșantion (de exemplu, prin selectarea în mod aleatoriu a unui eșantion dintr-o anumită colectivitate de interes, media calculată la nivelul acestuia este utilizată în vederea obținerii unei estimații a mediei populației generale).

În acest capitol, vom prezenta câteva aspecte importante ale procesului de cercetare statistică prin sondaj pornind de la estimarea punctuală a parametrilor unei anumite populații, respectiv utilizarea unei valori unice, calculată pe baza datelor din eșantion, în vederea obținerii unei estimări a valorii necunoscute a parametrului corespunzător din cadrul populației. O abordare mult mai consistentă din punct de vedere informațional constă însă în prezentarea unui interval de valori în care ne așteptăm, cu o anumită probabilitate sau nivel de încredere, să se afle valoarea parametrului populației studiate.

Estimarea punctuală a parametrului unei populații reprezintă o valoare unică a statisticii corespunzătoare, determinată la nivelul unui eșantion reprezentativ extras aleatoriu din aceasta.

De exemplu, un medic diabetolog dorește să estimeze valoarea medie a glicemiei à jeun (măsurată dimineața, pe nemâncate) la pacienții cu diabet zaharat de tip 2 dintr-o anumită zonă. Deoarece medicul nu poate măsura glicemia tuturor pacienților cu diabet de tip 2 din aria geografică deservită de spital, acesta va selecta în mod aleatoriu un eșantion de 50 de pacienți și va măsura glicemia acestora. Presupunem faptul că, în urma măsurătorilor, au fost obținute următoarele rezultate:

- Media eșantionului: $\bar{x} = 160$ mg/dL;
- Deviația standard a eșantionului est: $\sigma = 25$ mg/dL;
- Volumul eșantionului extras: $n = 50$ pacienți cu diabet zaharat.

Valoarea medie a glicemiei pacienților din eșantion $\bar{x} = 160$ mg/dL reprezintă „cea mai bună estimare" punctuală pentru glicemia medie μ din populația generală (totalul pacienților cu diabet zaharat de tip 2 din acea zonă). Această valoare poate fi folosită pentru a caracteriza nivelul general al glicemiei la toți pacienții cu diabet din acea arie geografică și poate constitui un etalon pentru diferite recomandări terapeutice sau pentru realizarea unor studii sau cercetări științifice ulterioare.

Un estimator punctual pentru valoarea unui parametru al unei populații nu ne oferă însă posibilitatea conturării unei imagini complete asupra fenomenului. Deși ne așteptăm ca valoarea estimatorului, calculată la nivel de eșantion, să fie apropiată de valoarea parametrului corespunzător din cadrul populației, în analiză dorim să cunoaștem în același timp și cât de aproape este acesta. Cu alte cuvinte, încercăm să cuantificăm incertitudinea care planează asupra estimării noastre punctuale sau nivelul de încredere atașat acestei presupuneri.

Scanned with OKEN Scanner

Astfel, ar fi mult mai interesant să putem afirma faptul că există 95% șanse ca valoarea medie „adevărată" a glicemiei tuturor pacienților cu diabet zaharat de tip 2 din acea zonă să fie cuprinsă între 152,88 mg/dL și 167,12 mg/dL. Sau, într-un mod mai general, 95% din intervalele de încredere construite prin utilizarea tuturor eșantioanelor de volum $n = 50$ extrase în mod aleatoriu din populația pacienților cu diabet zaharat de tip 2 din acea arie geografică vor conține valoarea adevărată a parametrului – „glicemia medie μ din populația generală".

În general, intervalul de încredere pentru media unei populații μ, calculat pe baza mediei eșantionului $\bar{x}$ și a erorii limită admisibile d este:

$$\bar{x} - d \leq \mu \leq \bar{x} + d \qquad (7.1.)$$

Eroarea limită admisibilă (d) reprezintă jumătate din lungimea intervalului de încredere în cazul unui indicator statistic estimat (de exemplu media populației μ), fiind valoarea maximă cu care estimarea punctuală (media eșantionului $\bar{x}$), poate devia de la valoarea reală a parametrului populației, pentru un anumit nivel de confidență stabilit în cadrul studiului. Eroarea limită admisibilă ne arată cât de precisă este estimarea mediei populației. Astfel, o valoare redusă a lui d indică o precizie mai bună a estimării, în timp ce o valoare ridicată a erorii limită este expresia unui grad de precizie mai scăzut.

În general, *intervalul de încredere* reprezintă o plajă de valori construită pe baza datelor din eșantion, în care este probabil să fie cuprinsă valoarea „necunoscută" și „adevărată" a parametrului corespunzător al populației generale (media, proporția, deviația standard, etc.), din care lotul de studiu a fost extras în mod aleatoriu.

Probabilitatea specifică estimării acestui interval se numește *nivel de încredere* și are de obicei în practică valoarea de 0,95 sau 95%. Aceasta exprimă "siguranța" cu care se afirmă faptul că intervalul de încredere construit cuprinde valoarea parametrului estimat, în cazul exemplului nostru, media populației generale μ:

$$P(\bar{x} - d \leq \mu \leq \bar{x} + d) = 1 - \alpha = 1 - 0{,}05 = 0{,}95 \qquad (7.2)$$

În cadrul relației 7.2, $\alpha = 1 - 0{,}95 = 0{,}05$ sau 5% reprezintă pragul sau nivelul de semnificație asociat estimării.

In concluzie, intervalul de încredere evidențiază nivelul de precizie al estimării parametrului din populația de referință, cu un anumit nivel de încredere.

Scanned with OKEN Scanner

7.2. INTERVALUL DE ÎNCREDERE PENTRU ESTIMAREA MEDIEI UNEI POPULAȚII

Metoda de lucru pornește generic de la ideea studierii variabilei aleatoare create prin utilizarea tuturor mediilor eșantioanelor de același volum extrase din aceeași populație generală

Conform teoremei limită centrală, în situația când volumul eșantionului este mare ($n \geq 30$) distribuția de sondaj a mediilor eșantioanelor poate fi aproximată printr-o distribuție normală având media μ și deviația standard $\sigma/\sqrt{n}$, indiferent de forma distribuției inițiale a valorilor variabilei studiate.

Chiar dacă *media populației generale nu este cunoscută*, distribuția de sondaj a mediilor eșantioanelor ne arată cum sunt repartizate valorile $\bar{x}$ în jurul lui μ. Astfel, aflăm informații despre toate diferențele $\bar{x} - \mu$ sau erorile de sondaj posibile.

Conform regulii empirice, există $1 - \alpha = 95\%$ șanse ca o anumită variantă x_i a unei variabile normal distribuite X, având media μ și deviația standard σ, să se situeze în intervalul:

$$\mu - 1{,}96 \cdot \sigma \leq x_i \leq \mu + 1{,}96 \cdot \sigma \tag{7.3}$$

Deoarece utilizăm în calcule distribuția de sondaj a mediilor eșantioanelor $\bar{x}$ în locul distribuției variantelor individuale ale variabilei x_i, vom folosi în locul deviației standard a populației generale σ, deviația standard a distribuției de sondaj a mediilor tuturor eșantioanelor de același volum n extrase în mod aleatoriu din aceasta sau eroarea standard a mediei $\sigma_{\bar{x}}$:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{7.4}$$

Rezultă imediat faptul că, în cadrul distribuției de sondaj a mediilor eșantioanelor există o probabilitate de 95% ca media eșantionului studiat $\bar{x}$ să se situeze între -1,96 și +1,96 erori standard $\sigma_{\bar{x}}$ de la media populației generale μ. Relația 7.3, transpusă în cazul distribuției de sondaj a mediilor, devine:

$$\mu - 1{,}96 \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + 1{,}96 \cdot \frac{\sigma}{\sqrt{n}} \tag{7.5}$$

Pentru a construi un interval de încredere $1 - \alpha = 95\%$ (nivel de semnificație $\alpha = 5\%$) pentru media populației generale μ pornind de la media eșantionului studiat $\bar{x}$ vom rearanja termenii din relația 7.5, pentru a exprima media populației μ în funcție de media eșantionului de lucru $\bar{x}$, astfel:

Scanned with OKEN Scanner

$$\bar{x} - 1{,}96 \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1{,}96 \cdot \frac{\sigma}{\sqrt{n}} \tag{7.6}$$

Pentru un nivel de semnificație $\alpha = 10\%$, intervalul de încredere $1 - \alpha = 90\%$ devine:

$$\bar{x} - 1{,}645 \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1{,}645 \cdot \frac{\sigma}{\sqrt{n}} \tag{7.7}$$

Pentru un nivel de semnificație $\alpha = 1\%$, intervalul de încredere $1 - \alpha = 99\%$ devine:

$$\bar{x} - 2{,}576 \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 2{,}576 \cdot \frac{\sigma}{\sqrt{n}} \tag{7.8}$$

Pentru a generaliza, intervalul de încredere pentru media unei populații va fi determinat prin transformarea distribuției de sondaj a mediilor eșantioanelor într-o distribuție normală standard, de medie zero și varianță 1. Considerăm o variabilă aleatoare repartizată normal $N(\mu, \sigma^2)$ pentru care dorim să estimăm intervalul de încredere pentru valoarea mediei (μ).

Dispunem de un eșantion de volum n extras aleatoriu din populația de volum N. Media datelor calculată la nivelul eșantionului este $\bar{x}$, iar media necunoscută a populației este μ. Vom utiliza în calcule o probabilitate de estimare de $1 - \alpha$ ceea ce înseamnă în cazul unei repartiții normale reduse $N(0,1)$ un *interval simetric* cuprins între $z_{\alpha_1} = -z_{1-\frac{\alpha}{2}}$ și $z_{\alpha_2} = +z_{1-\frac{\alpha}{2}}$:

$$P\left(-z_{1-\frac{\alpha}{2}} \leq z_{\bar{x}} \leq z_{1-\frac{\alpha}{2}}\right) = 1 - \alpha \tag{7.9}$$

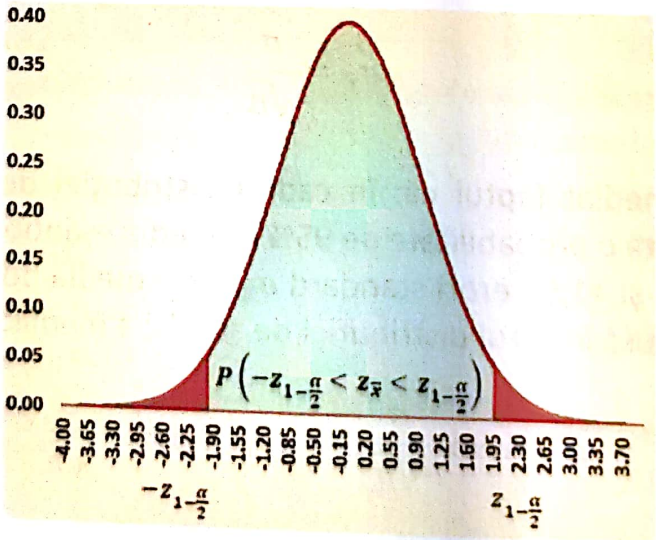


Figura 7.1. Probabilitatea de estimare $1 - \alpha$ în cadrul distribuției normale standard

Scanned with OKEN Scanner

$1-\alpha$ reprezintă coeficientul sau nivelul de încredere și $z_{1-\frac{\alpha}{2}}$ este acea valoare a lui z care delimitează o suprafață de $^{\alpha}/_{2}$ în extremitatea superioară a distribuției normale standard.

Valoarea $z_{\bar{x}}$ corespunzătoare mediei eșantionului $\bar{x}$ extras din populația generală cu media μ devine:

$$z_{\bar{x}} = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \tag{7.10}$$

Așa cum am văzut anterior, valoarea lui $z_{\bar{x}}$ este situată în intervalul simetric delimitat de valorile critice $z_{\alpha_1} = -z_{1-\frac{\alpha}{2}}$ și $z_{\alpha_2} = +z_{1-\frac{\alpha}{2}}$:

$$-z_{1-\frac{\alpha}{2}} \leq z_{\bar{x}} \leq z_{1-\frac{\alpha}{2}} \tag{7.11}$$

Înlocuind valoarea lui $z_{\bar{x}}$ din relația 7.10 în relația 7.11, obținem:

$$-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}} \tag{7.12}$$

Efectuând calculele în inegalitatea din relația 7.12 obținem intervalul de încredere pentru media populației generale estimat cu probabilitatea $1-\alpha$, în funcție de media eșantionului analizat ($\bar{x}$), deviația standard a populației generale (σ), volumul eșantionului (n) și valoarea critică a distribuției normale standard ($z_{1-\frac{\alpha}{2}}$) pentru un nivel de încredere de $1-\alpha$:

$$\bar{x} - z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \tag{7.13}$$

Sintetizând, putem afirma faptul că, există o probabilitate $1-\alpha$ ca intervalul $\bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ să conțină media adevărată a populației generale μ din care eșantionul de medie $\bar{x}$ a fost extras.

Pentru determinarea valorii critice $z_{1-\alpha/2}$ putem folosi tabelele distribuției normale standard sau putem genera direct, cu ajutorul programului Excel această valoare, utilizând de exemplu, pentru un nivel de încredere de 95% ($1-\alpha = 0{,}95$ și $\alpha = 0{,}05$) funcția *NORM.S.INV(probabilitate)*:

$$= NORM.S.INV(1-0{,}05/2) = NORM.S.INV(0{,}975) = 1{,}96$$

Intervalele de încredere pentru media unei populații μ au fost determinate anterior după formula 7.13 presupunând faptul că deviația standard a populației σ este cunoscută.

Scanned with OKEN Scanner

În cele mai multe situații practice însă, σ este necunoscută și vom utiliza ca estimator punctual al acesteia deviația standard ajustată, calculată la nivel de eșantion, prin raportarea sumei pătratelor abaterilor nivelurilor individuale de la media eșantionului la $n-1$ în loc de n, astfel:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{7.14}$$

Astfel, estimatorul deviației standard a distribuției de sondaj a mediilor eșantioanelor $s_{\bar{x}}$ devine:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{7.15}$$

În situația când deviația standard a populației generale σ nu este cunoscută, distribuția de sondaj a mediilor eșantioanelor urmează o distribuție probabilistică continuă *Student* (t), care este foarte apropiată de distribuția normală standard.

Distribuția Student prezintă următoarele caracteristici (figura 7.2):

- are o *formă de clopot*, fiind distribuită simetric in jurul mediei de valoare zero;
- are o *deviație standard mai mare decât 1;*
- există o *familie de distribuții t*, câte una pentru fiecare grad de libertate
- pe măsură ce volumul eșantionului n crește, distribuția t tinde către distribuția normală

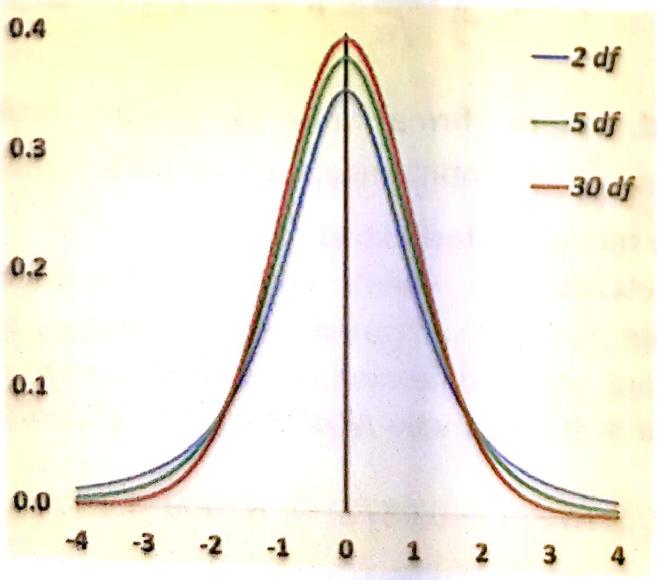


Figura 7.2. Distribuțiile Student cu 2, 5 și 30 de grade de libertate

Scanned with OKEN Scanner

În cele mai multe situații practice însă, σ este necunoscută și vom utiliza ca estimator punctual al acesteia deviația standard ajustată, calculată la nivel de eșantion, prin raportarea sumei pătratelor abaterilor nivelurilor individuale de la media eșantionului la $n-1$ în loc de n, astfel:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{7.14}$$

Astfel, estimatorul deviației standard a distribuției de sondaj a mediilor eșantioanelor $s_{\bar{x}}$ devine:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}} \tag{7.15}$$

În situația când deviația standard a populației generale σ nu este cunoscută, distribuția de sondaj a mediilor eșantioanelor urmează o distribuție probabilistică continuă *Student* (t), care este foarte apropiată de distribuția normală standard.

Distribuția Student prezintă următoarele caracteristici (figura 7.2):

- are o *formă de clopot*, fiind distribuită simetric in jurul mediei de valoare zero;
- are o *deviație standard mai mare decât 1;*
- există o *familie de distribuții t*, câte una pentru fiecare grad de libertate
- pe măsură ce volumul eșantionului n crește, distribuția t tinde către distribuția normală

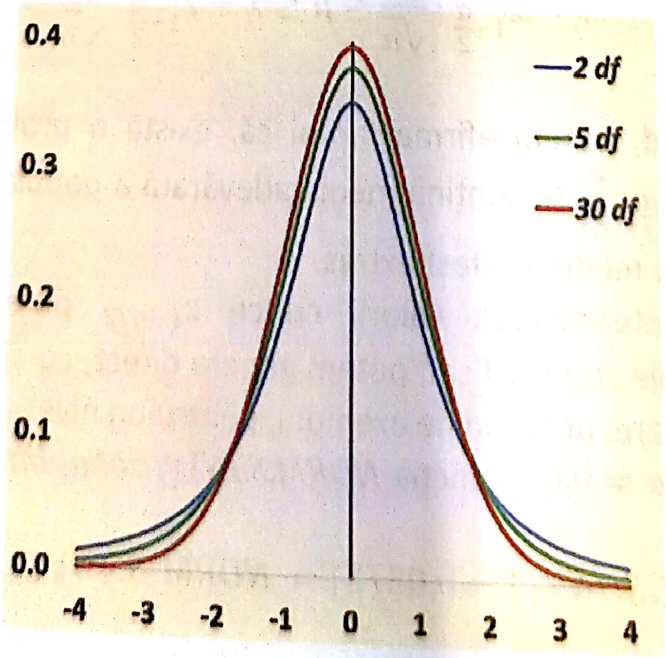


Figura 7.2. Distribuțiile Student cu 2, 5 și 30 de grade de libertate

Scanned with OKEN Scanner

Gradele de libertate sunt specifice variantelor x_i ale unei variabile statistice X care *au posibilitatea să varieze liber* în momentul în care *valoarea indicatorului statistic calculat pe baza lor a fost fixată*. În cazul mediei aritmetice avem $n-1$ grade de libertate. Astfel, pentru media unei grupe de 12 de studenți la un examen, care are valoarea 7,50 notele unui număr de $n-1=11$ studenți pot varia în mod liber, dar nota ultimului student luată în calculul acestui indicator statistic va fi „fixată" la valoarea necesară obținerii nivelului mediu de 7,50.

Având în vedere faptul că *distribuția Student (t) prezintă o împrăștiere mai mare decât distribuția normală redusă (z)*, valoarea absolută a lui t pentru un anumit nivel de încredere specificat este superioară valorii corespunzătoare a lui z. Astfel, pentru un nivel de încredere de 95%: $z_{0.975}=1{,}96$ și $t_{2,\ 0.975}=4{,}30$ (figura 7.3).

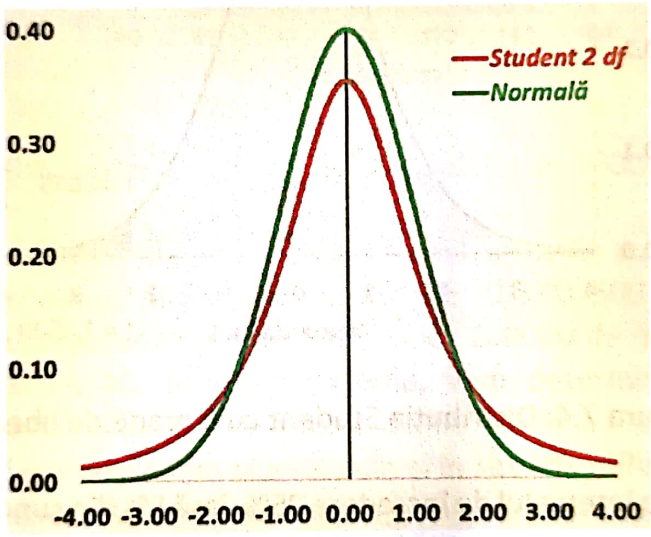


Figura 7.3. Distribuția Student cu 2 grade de libertate și distribuția Normală standard

Pe măsura utilizării unor eșantioane de volum superior ($n \geq 30$), diferența dintre intervalele de încredere generate prin utilizarea valorilor t și z devine neglijabilă; pentru $n \to \infty$, în cazul unui nivel de încredere de 95% ($1-\alpha=0{,}95$): $t_{n-1,\ 1-\frac{0{,}05}{2}} \approx z_{1-\frac{0{,}05}{2}} = 1{,}96$.

În concluzie, intervalul de încredere pentru media unei populații μ având σ necunoscută se obține astfel:

$$\bar{x} - t_{n-1,\ 1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1,\ 1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \qquad (7.16)$$

Scanned with OKEN Scanner

unde $t_{n-1,\ 1-\frac{\alpha}{2}}$ reprezintă valoarea lui t care delimitează o suprafață de $\alpha/2$ în extremitatea dreaptă a distribuției Student cu $n-1$ grade de libertate.

Pentru determinarea lui $t_{n-1,\ 1-\frac{\alpha}{2}}$ putem folosi tabelele distribuției Student sau putem genera direct cu ajutorul programului Excel această valoare, utilizând, de exemplu, pentru un nivel de încredere de 95% ($1-\alpha = 0{,}95$ și $\alpha = 0{,}05$) funcția *T.INV(probabilitate, grade de libertate)*:

$$= T.INV(1-0{,}05/2;9) = T.INV(0{,}975;9) = 2{,}2622.$$

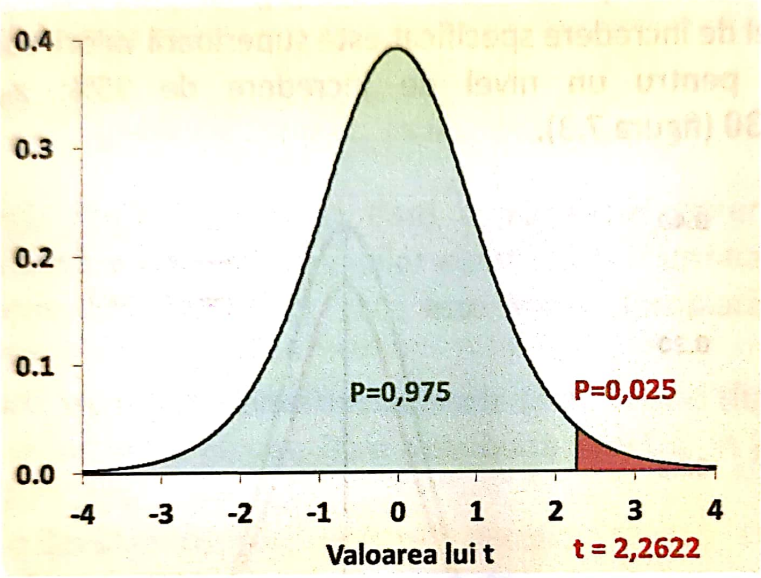


Figura 7.4. Distribuția Student cu 9 grade de libertate

Deoarece intervalul de încredere 95%, lasă 5% din suprafața de sub curba distribuției Student egal împărțită între cele 2 cozi laterale (2,5% în stânga și 2,5% în dreapta), avem nevoie de acea valoare a lui t în dreapta căreia să rămână 2,5% iar în partea stângă restul de 97,5% (figura 7.4).

Pentru exemplificare, considerăm distribuția fictivă a scorurilor IQ într-o populație „ipotetică" de 100000 persoane, având media de 100,13 și deviația standard 15,04. Scorurile IQ variază pe o plajă de valori cuprinsă între 38,60 și 170,18, fiind distribuite aproximativ normal.

Tabelul 7.1. Distribuția scorurilor IQ în populația de 100000 persoane

Persoana	Scorul IQ
1	76,11
2	113,33
...	...
100000	69,65

Scanned with OKEN Scanner

În continuare vom realiza o reprezentare grafică a datelor din tabelul 7.1, sub forma histogramei, pentru a evidenția mai clar forma distribuției normale a variabilei „scorul IQ". (figura 7.5).

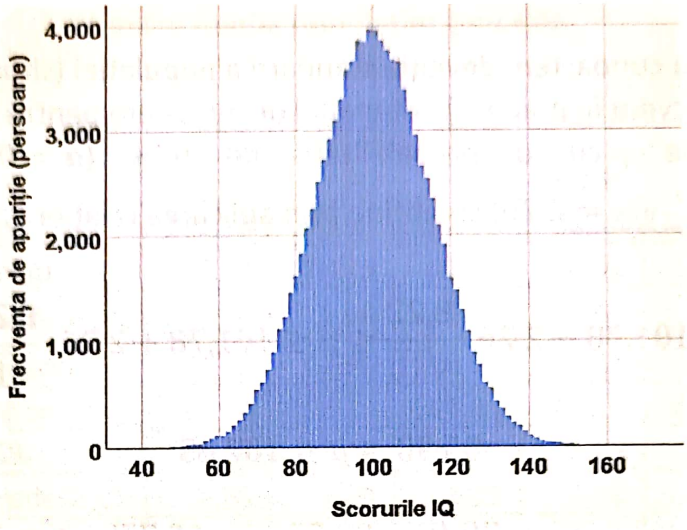


Figura 7.5. Histograma scorurilor IQ în populația de 100000 persoane

Vom calcula intervalele de încredere pentru media populației generale, prin utilizarea unor eșantioane de volume 10, 20, 30, 100, 500 și 1000 extrase în mod aleatoriu din colectivitatea generală a celor 100000 de persoane analizate prin prisma scorurilor IQ. În toate cazurile, vom determina intervalele de confidență pentru un nivel de încredere de 95% atât în ipoteza cunoașterii deviației standard σ a populației studiate cât și în situația utilizării, în calitate de estimator al acesteia, a deviației standard ajustate s, calculate la nivelul fiecărui eșantion în parte.

a) *Eșantion de volum* $n = 10$ ($\bar{x} = 101{,}78, s = 8{,}22$) extras în mod aleatoriu din populația generală de volum $N = 100000$ persoane ($\mu = 100{,}13, \sigma = 15{,}04$)

Tabelul 7.2. Scorurile IQ din eșantionul de volum $n = 10$ extras în mod aleatoriu din populația de 100000 persoane

100,20	107,37	94,93	102,90	100,97	97,36	102,80	87,29	105,36	118,58

- Dacă cunoaștem deviația standard a populației $\sigma = 15{,}04$ (situație care apare extrem de rar în practică), intervalul de încredere pentru media populației generale μ cu o probabilitate de 95% ($\alpha = 0{,}05$, $z_{1-\frac{\alpha}{2}} = z_{1-\frac{0{,}05}{2}} = z_{0{,}975} = 1{,}96$) se obține prin aplicarea relației 7.13:

Scanned with OKEN Scanner

$$101{,}78 - 1{,}96 \cdot \frac{15{,}04}{\sqrt{10}} \leq \mu \leq 101{,}78 + 1{,}96 \cdot \frac{15{,}04}{\sqrt{10}}$$

$$92{,}46 \leq \mu \leq 111{,}10$$

- Dacă nu cunoaștem deviația standard a populației (situațiile întâlnite cel mai frecvent în practică), intervalul de încredere pentru media populației generale μ cu o probabilitate de 95% ($\alpha = 0{,}05$, $t_{n-1,\ 1-\frac{\alpha}{2}} = t_{10-1,\ 1-\frac{0{,}05}{2}} = 2{,}26$) se obține prin aplicarea relației 7.16:

$$101{,}78 - 2{,}26 \cdot \frac{8{,}22}{\sqrt{10}} \leq \mu \leq 101{,}78 + 2{,}26 \cdot \frac{8{,}22}{\sqrt{10}}$$

$$95{,}90 \leq \mu \leq 107{,}65$$

b) *Eșantion de volum* $n = 20$ ($\bar{x} = 99{,}56$, $s = 10{,}75$) extras în mod aleatoriu din populația generală de volum $N = 100000$ persoane ($\mu = 100{,}13$, $\sigma = 15{,}04$)

Tabelul 7.3. Scorurile IQ din eșantionul de volum $n = 20$ extras în mod aleatoriu din populația de 100000 persoane

89,56	106,13	100,25	87,10	90,15	81,37	96,81	94,34	112,27	120,40
93,51	112,25	103,01	104,90	106,89	88,17	91,77	96,69	97,23	118,31

- Dacă cunoaștem deviația standard a populației, IC 95% (relația 7.13):

$$99{,}56 - 1{,}96 \cdot \frac{15{,}04}{\sqrt{20}} \leq \mu \leq 99{,}56 + 1{,}96 \cdot \frac{15{,}04}{\sqrt{20}}$$

$$92{,}96 \leq \mu \leq 106{,}15$$

- Dacă nu cunoaștem deviația standard a populației, IC 95% (relația 7.16):

$$99{,}56 - 2{,}09 \cdot \frac{10{,}75}{\sqrt{20}} \leq \mu \leq 99{,}56 + 2{,}09 \cdot \frac{10{,}75}{\sqrt{20}}$$

$$94{,}52 \leq \mu \leq 104{,}59$$

Continuăm calculul intervalelor de confidență pentru un nivel de încredere de 95% în mod similar în cazul eșantioanelor de volum 30, 100, 500 și

Scanned with OKEN Scanner

1000 de persoane, extrase în mod aleatoriu din aceeași populație generală, rezultatele fiind centralizate în următorul tabel:

Tabelul 7.4. Determinarea intervalelor de încredere pentru media populației generale

Eșantion	Nr. 1	Nr. 2	Nr. 3	Nr. 4	Nr. 5	Nr. 6
Volum eșantion (n)	*10*	*20*	*30*	*100*	*500*	*1000*
Medie eșantion ($\bar{x}$)	101,78	99,56	102,08	98,16	100,75	99,66
Deviație std. eșantion (s)	8,22	10,75	12,38	16,01	14,77	15,27
Coeficient de încredere (z)	1,96	1,96	1,96	1,96	1,96	1,96
IC 95% limita inferioară	92,46	92,96	96,70	95,21	99,43	98,73
IC 95% limita superioară	111,10	106,15	107,46	101,10	102,06	100,59
Dimensiune IC 95%	18,64	13,18	10,76	5,89	2,64	1,86
Coeficient de încredere (t)	2,26	2,09	2,05	1,98	1,96	1,96
IC 95% limita inferioară	95,90	94,52	97,46	94,98	99,45	98,71
IC 95% limita superioară	107,65	104,59	106,70	101,33	102,04	100,61
Dimensiune IC 95%	11,76	10,06	9,24	6,35	2,59	1,89

Analizând informațiile prezentate în tabelul 7.4 se remarcă faptul că, pe măsura creșterii volumului eșantionului de studiu n, eroarea standard a mediei ($\sigma/\sqrt{n}$ sau $s/\sqrt{n}$, după caz) se reduce. De asemenea, lucrând cu eșantioane de volum superior, dimensiunea intervalelor de confidență, în care sperăm că se va situa media populației generale, se micșorează corespunzător, conducând în mod evident la o creștere a preciziei estimării. În același timp, observăm că în situația eșantioanelor de volum mare ($n \geq 30$), diferența dintre intervalele de încredere generate prin utilizarea coeficienților t și z devine neglijabilă.

Interpretarea probabilistică a intervalelor de încredere: dacă replicăm de un număr foarte mare de ori experimentul de extragere a eșantioanelor de volum n dintr-o populație normal distribuită de volum N, având o deviație standard cunoscută în prealabil (σ) sau necunoscută (s), $100(1-\alpha)$ procente din toate intervalele de forma $\bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ sau $\bar{x} \pm t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$, după caz, vor include media aritmetică adevărată a populației generale (μ). De obicei, $100(1-\alpha) = 95\%$ dintre toate intervalele de încredere construite în această manieră vor conține în interiorul lor media populației generale (μ).

Interpretarea practică a intervalelor de încredere: atunci când eșantionul a fost extras în mod aleatoriu dintr-o populație normal distribuită, având o deviație standard cunoscută în prealabil (σ) sau necunoscută (s), apreciem cu un nivel de confidență de $100(1-\alpha)$ procente, de obicei 95%, că intervalul de forma

Scanned with OKEN Scanner

$\bar{x} \pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ sau $\bar{x} \pm t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$, după caz, va include media aritmetică adevărată a populației generale (μ).

În toate cazurile prezentate anterior, riscul α (de obicei 5%) a fost distribuit bilateral și simetric ($\alpha/2 = 2{,}5\%$ în stânga și $\alpha/2 = 2{,}5\%$ în dreapta). În funcție de a abordarea cercetării putem avea și intervale de încredere unilaterale (la stânga sau la dreapta), situații în care riscul este distribuit integral doar într-o singură parte.

Eroarea limită admisibilă $d = z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ sau $d = t_{n-1,\ 1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$ poate fi calculată cu ajutorul programului Excel utilizând funcțiile:

$$CONFIDENCE.NORM(\alpha, \sigma, n) \quad \text{sau} \quad CONFIDENCE.T(\alpha, s, n).$$

Continuând exemplul practic prezentat anterior, vom extrage aleatoriu 50 de eșantioane de volum $n = 50$ din populația generală $N = 100000$ persoane și vom construi 50 de intervale de încredere pentru media populației (μ), utilizând drept estimator al deviației standard a populației studiate deviația standard ajustată (s), determinată la nivelul fiecărui eșantion (distribuția Student).

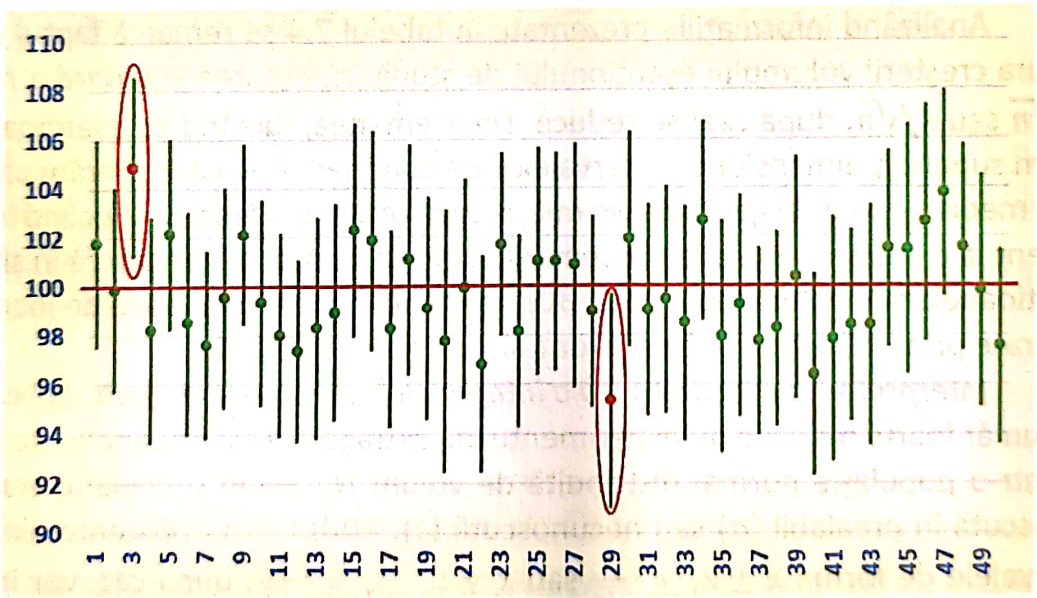


Figura 7.6. Intervalele de încredere 95% pentru media populației construite în baza a 50 de eșantioane de volum $n = 50$ extrase în mod aleatoriu din populația de 100000 persoane

Dintre cele 50 de intervale de încredere 95% create pe baza eșantioanelor de volum $n = 50$ extrase în mod aleatoriu din populația de 100000 persoane, 48 conțin media populației generale ($\mu = 100$), excepție făcând două dintre acestea: nr. 3 (101,21; 108,59) și nr. 29 (90,98; 99,68).

Scanned with OKEN Scanner

7.3. INTERVALUL DE ÎNCREDERE PENTRU ESTIMAREA DIFERENȚEI DINTRE MEDIILE A DOUĂ POPULAȚII

Uneori apare în practică necesitatea estimării diferenței între mediile a două populații distribuite normal μ_1 și μ_2, din cadrul cărora a fost extras în mod aleatoriu câte un eșantion de volum n_1 și n_2, având mediile $\bar{x}_1$ și $\bar{x}_2$.

Ca și în cazul mediei aritmetice, vom construi distribuția de sondaj a diferenței mediilor tuturor eșantioanelor de volum n, luând în calcul toate perechile posibile de 2 medii determinate la nivel de eșantion. Aceste diferențe ale mediilor eșantioanelor $\bar{x}_1 - \bar{x}_2$ vor fi grupate în funcție de frecvența lor de apariție, fiind obținută în acest mod o distribuție normală, cu *media*:

$$\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2 \tag{7.17}$$

și *varianța*:

$$\sigma^2_{\bar{x}_1-\bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \tag{7.18}$$

Deviația standard a acestei distribuții sau eroarea standard a diferenței mediilor eșantioanelor va fi:

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{7.19}$$

Similar cu situația intervalului de încredere pentru estimarea mediei unei populații, $(\bar{x}_1 - \bar{x}_2)$ reprezintă un estimator nedeplasat al diferenței mediilor celor 2 populații $(\mu_1 - \mu_2)$, având varianța $\sigma^2_{\bar{x}_1-\bar{x}_2}$.

Atunci când variațele sau dispersiile populațiilor (σ_1^2 și σ_2^2) sunt cunoscute, intervalul de încredere pentru $\mu_1 - \mu_2$ garantat cu probabilitatea de $100(1-\alpha)$ este:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \tag{7.20}$$

Simpla examinare a intervalului de încredere pentru diferența dintre mediile celor 2 populații studiate ne oferă suficientă informație pentru a putea decide dacă acestea pot fi sau nu considerate egale. Atunci când intervalul de încredere *nu include valoarea zero*, putem afirma faptul că acesta oferă dovezi în sprijinul ideii că mediile celor 2 populații nu sunt egale. În situația contrară, când

Scanned with OKEN Scanner

intervalul conține valoarea zero, concluzionăm că mediile populațiilor analizate pot fi egale.

Pentru a exemplifica, presupunem că avem două populații de persoane, dintre care una a fost expusă unor factori care sunt asociați retardului mintal (infecții, anomalii cromozomiale, anomalii genetice, cauze nutriționale, toxice, traumatisme, etc.), iar cea de a doua nu a experimentat aceste condiții. Distribuția scorurilor de inteligență (IQ) în cele două populații se presupune că este aproximativ normală, cu o deviație standard de 21 și respectiv 16. Extragem câte un eșantion din fiecare populație ($n_1 = 15,\ n_2 = 20$) și determinăm pentru fiecare scorul mediu de inteligență, respectiv $\bar{x}_1 = 83$ și $\bar{x}_2 = 101$. Vom construi intervalul de încredere estimat cu o probabilitate de 95% pentru diferența mediilor celor 2 populații.

Utilizăm în calcule drept estimator punctual nedeplasat al diferenței mediilor celor 2 populații ($\mu_1 - \mu_2$):

$$\bar{x}_1 - \bar{x}_2 = 83 - 101 = -18$$

Coeficientul de încredere corespunzător unei probabilități de 95% este:

$$z_{1-\frac{\alpha}{2}} = 1{,}96$$

Eroarea standard a diferențelor dintre mediile eșantioanelor:

$$\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{21^2}{15} + \frac{16^2}{20}} = 6{,}50$$

Intervalul de încredere estimat cu o probabilitate de 95% pentru diferența mediilor celor 2 populații:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = -18 \pm 1{,}96 \cdot 6{,}50$$

$$-30{,}74 \leq \mu_1 - \mu_2 \leq -5{,}26$$

Astfel, suntem în proporție de 95% „convinși" de faptul că diferența reală dintre mediile celor 2 populații se află cuprinsă în interiorul intervalului $(-30{,}74;\ -5{,}26)$, deoarece repetând procesul de eșantionare, 95% dintre intervalele construite în această manieră vor include diferența dintre mediile reale ale populațiilor. Întrucât intervalul de încredere determinat anterior nu conține valoarea zero, concluzionăm că *mediile celor două populații nu sunt*

Scanned with OKEN Scanner

egale. Cu alte cuvinte, scorurile medii de inteligență din cele două populații comparate nu sunt egale sau „diferențele dintre acestea sunt semnificative statistic" și nu au apărut întâmplător.

Determinarea intervalelor de încredere pentru diferența mediilor a 2 populații care nu urmează o distribuție normală se realizează într-o manieră similară în situația când volumele ambelor eșantioane n_1 și n_2 sunt mai mari de 30 de unități (conform teoremei limită centrală). Dacă variațele celor 2 populații (σ_1^2 și σ_2^2) nu sunt cunoscute, în calcule vom folosi valorile calculate la nivelul fiecărui eșantion (s_1^2 și s_2^2) pentru a le estima.

Similar cu estimarea intervalului de încredere pentru media unei populații, în situația când deviațiile standard ale celor 2 populații (σ_1^2 și σ_2^2) nu sunt cunoscute distribuția de sondaj a diferenței mediilor eșantioanelor urmează o repartiție probabilistică continuă Student (t), care este foarte apropiată de distribuția normală standard. Utilizarea distribuției „t" se bazează pe faptul că distribuțiile celor 2 populații studiate sunt normale.

În practică, deosebim 2 situații distincte:

a) *variațele celor două populații sunt egale*

În situația când putem considera că variațele celor 2 populații studiate sunt egale, variațele ajustate, calculate la nivelul fiecărui eșantion, pot fi considerate drept estimări ale variației comune. Pentru a obține un *estimator „combinat" al variației comune* vom calcula media aritmetică a variațelor celor două eșantioane, ponderate cu gradele de libertate ale fiecăruia $(n_1 - 1)$, respectiv $(n_2 - 1)$:

$$s_p^2 = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{(n_1 - 1) + (n_2 - 1)} \tag{7.21}$$

Eroarea standard a diferențelor dintre mediile eșantioanelor:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} \tag{7.22}$$

Intervalul de încredere pentru $\mu_1 - \mu_2$ garantat cu probabilitatea de $100(1 - \alpha)$ este:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2,\ 1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} \tag{7.23}$$

în care, $t_{n_1+n_2-2,\ 1-\frac{\alpha}{2}}$ reprezintă valoarea critică a testului Student pentru un nivel de încredere $1 - \alpha$ și $n_1 + n_2 - 2$ grade de libertate.

Scanned with OKEN Scanner

Considerăm cu titlu de exemplu compararea eficacității a două medicamente antihipertensive A și B. Sunt selectate două grupuri de pacienți care au hipertensiune, pacienții din primul grup fiind tratați cu medicamentul A, iar cei din al doilea grup cu medicamentul B. După 8 săptămâni, se măsoară reducerea medie a tensiunii arteriale sistolice (TAS) pentru fiecare dintre grupuri.

- *grupul A* ($n_1 = 30$ pacienți) are o reducere medie a TAS $\bar{x}_1 = 15$ mmHg, cu o deviație standard $s_1 = 5$ mmHg;
- *grupul B* ($n_2 = 30$ *pacienți*) are o reducere medie a TAS $\bar{x}_2 = 12$ mmHg, cu o deviație standard de $s_2 = 5$ mmHg.

Presupunem faptul că variațele celor două populații din care provin eșantioanele studiate sunt egale (${\sigma_1}^2 = {\sigma_2}^2$), sau nivelul variabilității datelor din cele două populații studiate este relativ similar.

Estimatorul „combinat" al varianței comune se determină utilizând relația (7.21):

$$s_p^2 = \frac{(30-1)\cdot 5^2 + (30-1)\cdot 5^2}{(30-1)+(30-1)} = 25$$

Citim din tabelele distribuției Student valoarea critică a testului t pentru un nivel de încredere $1 - \alpha = 0{,}95$ și $n_1 + n_2 - 2 = 30 + 30 - 2 = 58$ grade de libertate sau utilizăm formula $= T.INV(0{,}975; 58)$ din programul Excel și obținem:

$$t_{58,\ 0{,}975} = 2{,}0017$$

Conform relației 7.23, *intervalul de încredere* pentru diferența mediilor din cele două populații $\mu_1 - \mu_2$, garantat cu probabilitatea de $100(1-\alpha) = 95\%$, este:

$$(15-12) \pm 2{,}0017 \cdot \sqrt{\frac{25}{30} + \frac{25}{30}} = 3 \pm 2{,}58$$

Informațiile pot fi prezentate și în manieră grafică, reprezentând vizual intervalul de încredere estimat cu o probabilitate de 95%, respectiv $(3 \pm 2{,}58)$ mmHg pentru diferența mediilor celor două populații de pacienți hipertensivi, care au fost tratați cu medicamentele A și B și evidențiind în cadrul figurii apartenența sau nu a valorii zero la acest interval (figura 7.7).

Scanned with OKEN Scanner

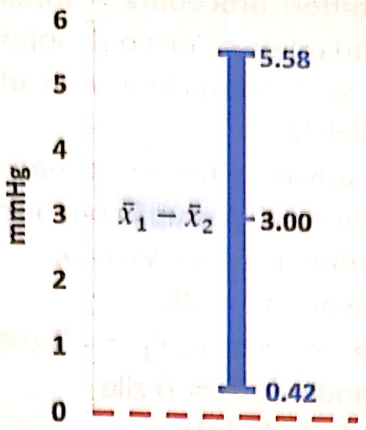


Figura 7.7. Intervalul de încredere 95% pentru diferența dintre reducerea medie a tensiunii arteriale în cazul celor 2 medicamente

Considerând un nivel de încredere de 95%, diferența dintre reducerea medie a TAS pentru cele două medicamente (A și B) se situează între 0,42 mmHg și 5,58 mmHg (figura 7.7). Deoarece intervalul nu include valoarea zero, putem concluziona faptul că există o diferență semnificativă statistic între eficacitatea acestor două medicamente, medicamentul A având un efect mai puternic asupra reducerii tensiunii arteriale comparativ cu medicamentul B.

b) *variațele celor două populații nu sunt egale*

Dacă nu putem considera faptul că cele 2 varianțe ale populațiilor analizate sunt egale, intervalul de încredere pentru $\mu_1 - \mu_2$ garantat cu probabilitatea de $100(1-\alpha)$ este:

$$(\bar{x}_1 - \bar{x}_2) \pm t'_{df,\ 1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \tag{7.24}$$

în care vom utiliza valoarea critică a testului Student $t'_{df,\ 1-\frac{\alpha}{2}}$ pentru un nivel de încredere $1 - \alpha$ și un număr efectiv de grade de libertate df, determinat prin formula Welch:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_1}\right)^2}{n_2 - 1}} \tag{7.25}$$

Pentru o mai bună înțelegere a noțiunilor, considerăm un exemplu referitor la compararea timpului de recuperare după două tipuri de intervenții chirurgicale. Se dorește compararea timpului mediu de recuperare (în zile) pentru pacienții care au suferit două tipuri de intervenții chirurgicale în vederea reparării

Scanned with OKEN Scanner

ligamentului încrucișat anterior: procedura A (minim invazivă) și procedura B (clasică). Presupunem că varianțele celor două populații din care au fost extrase eșantioanele sunt inegale și în consecință vom utiliza testul Student pentru varianțe inegale (metoda Welch).

Datele centralizate privind timpii de recuperare, pentru cele două tipuri de intervenții chirurgicale, se prezintă după cum urmează:

Eșantionul A (procedura minim invazivă):

- Număr pacienți: $n_1 = 20$
- Medie timp recuperare: $\bar{x}_1 = 25$ zile
- Deviație standard: $s_1 = 6$ zile

Eșantionul B (procedura clasică):

- Număr pacienți: $n_2 = 22$
- Medie timp recuperare: $\bar{x}_2 = 30$ zile
- Deviație standard: $s_2 = 8$ zile

Vom construi un *interval de încredere de 95%* pentru diferența dintre mediile celor două populații.

Calculăm numărul de grade de libertate prin metoda Welch (formula 7.25):

$$df = \frac{\left(\frac{6^2}{20} + \frac{8^2}{22}\right)^2}{\frac{\left(\frac{6^2}{20}\right)^2}{20-1} + \frac{\left(\frac{8^2}{22}\right)^2}{22-1}} \approx 38{,}7$$

Pentru un interval de încredere de 95%, valoarea critică a testului t' pentru un nivel de încredere $1 - \alpha = 0{,}95$ și $df = 38$ grade de libertate:

$$t_{38,\ 0{,}975} = 2{,}0244$$

Conform relației 7.24, *intervalul de încredere* pentru diferența mediilor din cele două populații $\mu_1 - \mu_2$, garantat cu probabilitatea de $100(1-\alpha) = 95\%$, este:

$$(25 - 30) \pm 2{,}0244 \cdot \sqrt{\frac{6^2}{20} + \frac{8^2}{22}} = -5 \pm 4{,}39$$

Realizăm și în acest caz reprezentarea grafică a intervalului de încredere pentru diferența timpului mediu de recuperare pentru cele două tipuri de intervenții chirurgicale, minim invazivă și clasică (figura 7.8).

Scanned with OKEN Scanner

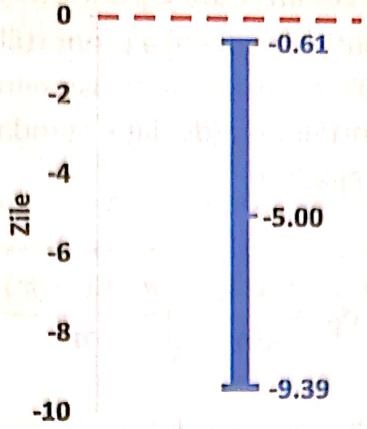


Figura 7.8. Intervalul de încredere 95% pentru diferența dintre timpii medii de recuperare ai pacienților care au suferit cele două tipuri de intervenții chirurgicale

Pentru un nivel de încredere de 95%, diferența medie dintre timpii de recuperare în cazul procedurilor A (minim invazivă) și B (clasică) este cuprinsă între $-9{,}39$ zile și $-0{,}61$ zile. Deoarece intervalul nu include valoarea zero, putem concluziona că există o diferență semnificativă statistic între cele două tipuri de intervenții chirurgicale, procedura A (minim invazivă) reducând în mod evident timpul de recuperare al pacienților în comparație cu procedura B (clasică).

7.4. INTERVALUL DE ÎNCREDERE PENTRU ESTIMAREA PROPORȚIEI UNEI POPULAȚII

Alături de estimarea mediei, un loc important în analiza statistică este deținut de studiul proporțiilor (ponderilor) deținute de diferite categorii/grupe în ansamblul populației de referință (proporția persoanelor de sex masculin dintr-o localitate, proporția pacienților care s-au vindecat în urma aplicării unui tratament, ponderea persoanelor imune la o anumită boală dintr-o zonă, etc.).

Estimarea proporției π a unei populații în funcție de proporția p a unui eșantion reprezentativ extras din aceasta se realizează într-o manieră similară cu estimarea mediei, proporția putând fi asimilată unui caz particular al mediei aritmetice (subcapitolul 6.5), întâlnit în situația unei variabile alternative (binare), care poate lua doar 2 valori (1 – în cazul variantelor variabilei care posedă caracteristica și 0 în rest).

În situația când volumul eșantioanelor este suficient de mare ($n \cdot p \geq 5$ și $n \cdot (1-p) \geq 5$) distribuția de sondaj a tuturor proporțiilor eșantioanelor de volum n extrase dintr-o populație de volum N tinde către repartiția normală.

Scanned with OKEN Scanner

Dacă populația generală de volum N are o proporție π și o deviație standard $\sigma = \sqrt{\pi \cdot (1-\pi)}$, atunci distribuția de sondaj a proporțiilor tuturor eșantioanelor de volum n extrase aleatoriu din aceasta are de asemenea o medie a proporțiilor π și o eroare standard a proporțiilor σ_p (deviație standard a distribuției de sondaj a proporțiilor eșantioanelor) egală cu:

$$\sigma_p = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\pi \cdot (1-\pi)}{n}} \tag{7.26}$$

În vederea estimării parametrului π, proporția din populația generală, vom utiliza în locul lui ca estimator punctual proporția p a eșantionului extras aleatoriu din cadrul populației. Astfel, relația 7.26 devine:

$$\sigma_p = \sqrt{\frac{p \cdot (1-p)}{n}} \tag{7.27}$$

Distribuția de sondaj a proporțiilor eșantioanelor poate fi convertită într-o distribuție normală standard prin schimbarea de variabilă:

$$z_p = \frac{p-\pi}{\sigma_p} = \frac{p-\pi}{\sigma/\sqrt{n}} = \frac{p-\pi}{\sqrt{\frac{p \cdot (1-p)}{n}}} \tag{7.28}$$

Intervalul de încredere pentru proporția populației generale π, garantat cu probabilitatea de $100(1-\alpha)$, de obicei 95%, se obține aplicând formula generală:

$$P\left(-z_{1-\frac{\alpha}{2}} < z_p < z_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$1-\alpha$ reprezintă coeficientul sau nivelul de încredere și $z_{1-\frac{\alpha}{2}}$ este acea valoare a lui z care delimitează o suprafață de $\alpha/2$ în extremitatea superioară a distribuției normale standard.

Valoarea lui z_p este situată în intervalul simetric delimitat de valorile critice $z_{\alpha_1} = -z_{1-\frac{\alpha}{2}}$ și $z_{\alpha_2} = +z_{1-\frac{\alpha}{2}}$:

$$-z_{1-\frac{\alpha}{2}} \leq z_p \leq z_{1-\frac{\alpha}{2}}. \tag{7.29}$$

Înlocuind valoarea lui z_p din relația 7.28 în relația 7.29, obținem:

Scanned with OKEN Scanner

$$-z_{1-\frac{\alpha}{2}} \leq \frac{p-\pi}{\sqrt{\frac{p\cdot(1-p)}{n}}} \leq z_{1-\frac{\alpha}{2}} \tag{7.30}$$

Efectuând calculele în inegalitatea din relația 7.30 obținem intervalul de încredere pentru proporția populației generale estimat cu probabilitatea $1-\alpha$, în funcție de proporția eșantionului analizat (p), volumul eșantionului (n) și valoarea critică a distribuției normale standard ($z_{1-\frac{\alpha}{2}}$) pentru un nivel de încredere de $1-\alpha$:

$$p - z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p\cdot(1-p)}{n}} \leq \pi \leq p + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p\cdot(1-p)}{n}} \tag{7.31}$$

De obicei, în practică utilizăm un nivel de semnificație $\alpha = 0{,}05$ (sau 5%) simetric, în care $z_{1-\frac{\alpha}{2}} = 1{,}96$ reprezintă abscisa corespunzătoare probabilității cumulate: $P = 1-\frac{\alpha}{2} = 1-\frac{0{,}05}{2} = 1-0{,}025 = 0{,}975$

Sintetizând, putem afirma faptul că, există o probabilitate $1-\alpha$ ca intervalul $p \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p\cdot(1-p)}{n}}$ să conțină proporția adevărată a populației generale π din care eșantionul de proporție p a fost extras.

Pentru a exemplifica din punct de vedere practic determinarea intervalului de încredere vom încerca să realizăm o estimare a proporției pacienților care dezvoltă o reacție alergică ușoară după administrarea unui nou vaccin. Presupunem că a fost selectat în mod aleatoriu un eșantion de volum $n = 1200$ de pacienți vaccinați, dintre care, un număr de 36 de pacienți raportează reacții alergice ușoare (erupții cutanate, mâncărimi, roșeață).

Vom utiliza proporția p calculată la nivel de eșantion (ponderea pacienților care au raportat reacții alergice ușoare la administrarea vaccinului din eșantion) în calitate de estimator punctual al parametrului π din populația generală:

$$p = \frac{36}{1200} = 0{,}03$$

Înlocuim datele în formula 7.31 și obținem intervalul de încredere 95% pentru proporția pacienților cu reacții alergice la vaccin din populația generală:

$$0{,}03 - 1{,}96 \cdot \sqrt{\frac{0{,}03\cdot(1-0{,}03)}{1200}} < \pi < 0{,}03 + 1{,}96 \cdot \sqrt{\frac{0{,}03\cdot(1-0{,}03)}{1200}}$$

Scanned with OKEN Scanner

Prin efectuarea calculelor, obținem intervalul de încredere în care cu o probabilitate de 95% se va situa proporția din cadrul populației studiate:

$$0{,}0203 < \pi < 0{,}0397$$

În consecință, putem afirma, cu un nivel de încredere de 95%, faptul că proporția reală a pacienților din populația generală, din care provine eșantionul analizat, care prezintă reacții alergice ușoare după administrarea respectivului vaccin este cuprinsă între 2,03% și 3,97%. Sau, din punct de vedere probabilistic, în cadrul procesului de eșantionare repetată, aproximativ 95% din intervalele construite în acest mod (pentru toate eșantioanele de același volum $n = 1200$) vor conține proporția reală existentă în populația generală.

Acest tip de analiză statistică este foarte util în testele clinice pentru a evalua siguranța și eficacitatea unui nou tratament sau vaccin. Astfel, în situația în care proporția reacțiilor alergice ușoare este considerată acceptabilă de către autoritățile de specialitate, vaccinul poate fi aprobat pentru utilizare. De asemenea, atunci când procentul de reacții alergice ar fi mai ridicat decât un nivel prag considerat sigur, ar fi necesare investigații suplimentare sau chiar ajustări ale compoziției vaccinului. Nu în ultimul rând, personalul medical de specialitate poate folosi aceste date statistice în scopul informării pacienților cu privire la riscul reacțiilor alergice ușoare în administrarea vaccinului.

7.5. INTERVALUL DE ÎNCREDERE PENTRU ESTIMAREA DIFERENȚEI DINTRE PROPORȚIILE A DOUĂ POPULAȚII

Evidențierea magnitudinii diferențelor dintre proporțiile a două populații este în foarte multe cazuri elementul central al unei cercetări. Putem astfel efectua comparații între 2 grupe de vârstă, 2 grupuri socio-economice, 2 grupuri de pacienți diagnosticați în funcție de proporția pe care o posedă o anumită caracteristică de interes în studiu.

Un estimator nedeplasat al diferenței dintre proporțiile a două populații $(\pi_1 - \pi_2)$ este reprezentat diferența dintre proporțiile a două eșantioane extrase în mod aleatoriu din fiecare dintre acestea $(p_1 - p_2)$. După cum a fost prezentat în capitolul anterior, în situația când volumele eșantioanelor sunt suficient de mari $n_1 \cdot p_1 \geq 5, n_1 \cdot (1 - p_1) \geq 5, n_2 \cdot p_2 \geq 5$ și $n_2 \cdot (1 - p_2) \geq 5$ și proporțiile celor două populații nu sunt foarte aproape de 0 sau 1, se aplică teorema limită centrală, astfel încât putem utiliza și în acest caz distribuția normală pentru a construi intervalele de încredere.

Eroarea standard a proporțiilor $\sigma_{p_1 - p_2}$ (deviația standard a distribuției de sondaj a diferenței dintre proporțiile eșantioanelor) este:

Scanned with OKEN Scanner

$$\sigma_{p_1-p_2} = \sqrt{\frac{\sigma_{p_1}^2}{n_1} + \frac{\sigma_{p_2}^2}{n_2}} = \sqrt{\frac{p_1 \cdot (1-p_1)}{n_1} + \frac{p_2 \cdot (1-p_2)}{n_2}} \quad (7.32)$$

Intervalul de încredere pentru diferența dintre proporțiile a două populații ($\pi_1 - \pi_2$), estimat cu probabilitatea $1-\alpha$, în funcție de proporțiile celor două eșantioane analizate (p_1 și p_2), volumele eșantioanelor (n_1 și n_2) și valoarea critică a distribuției normale standard ($z_{1-\frac{\alpha}{2}}$) pentru un nivel de încredere de $1-\alpha$:

$$(p_1 - p_2) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p_1 \cdot (1-p_1)}{n_1} + \frac{p_2 \cdot (1-p_2)}{n_2}} \quad (7.33)$$

Exemplu: un colectiv de cercetători dorește să compare eficacitatea a două metode diferite de renunțare la fumat: terapie de substituție nicotinică (plasturi, pastile sau gumă) vs. consiliere psihologică. Studiul demarează prin selectarea în mod aleatoriu a două grupuri de persoane fumătoare care vor fi urmărite pe parcursul unei perioade de 6 luni, după cum urmează:

- Grupul 1: Terapie de substituție nicotinică ($n_1 = 400$ participanți)
  - După 6 luni s-au lăsat de fumat 112 persoane
  - Calculăm proporția grupului 1: $p_1 = 112/400 = 0{,}28$
- Grupul al 2-lea: Consiliere psihologică ($n_2 = 350$ participanți)
  - După 6 luni s-au lăsat de fumat 81 persoane
  - Calculăm proporția grupului al 2-lea: $p_2 = 85/350 = 0{,}24$

La prima vedere, realizând doar o analiză descriptivă a eficacității celor două metode de tratament la nivelul fiecărui eșantion, observăm faptul că proporția persoanelor care s-au lăsat de fumat după cele 6 luni de tratament este mai mare în grupul 1 ($p_1 = 0{,}28$), care a urmat terapia de substituție nicotinică comparativ cu grupul 2 ($p_2 = 0{,}24$), care s-a bazat pe consilierea psihologică.

Determinăm intervalul de încredere 95% ($z_{1-\frac{\alpha}{2}} = 1{,}96$) pentru diferența dintre proporțiile celor două populații din care provin eșantioanele noastre, utilizând relația 7.33:

$$(0{,}28 - 0{,}24) \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{0{,}28 \cdot (1-0{,}28)}{400} + \frac{0{,}24 \cdot (1-0{,}24)}{350}}$$

După efectuarea calculelor obținem:

$$-0{,}026 \leq \pi_1 - \pi_2 \leq 0{,}100$$

Scanned with OKEN Scanner

În continuare vom realiza și o reprezentare grafică a diferenței dintre proporțiile celor care au renunțat la fumat, exprimată procentual:

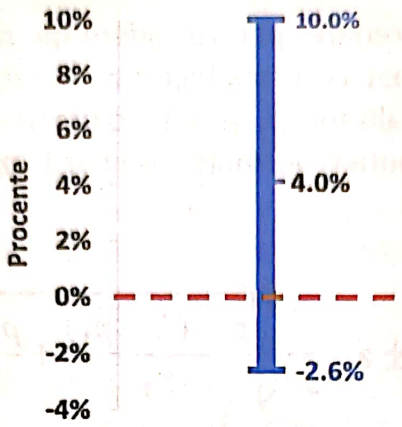


Figura 7.9. Intervalul de încredere 95% pentru diferența dintre proporțiile celor care au renunțat la fumat în cadrul celor două tipuri de terapie

Analizând graficul din figura 7.9, observăm că intervalul de încredere 95% construit pentru diferența proporțiilor celor două populații, din care au fost extrase în mod aleatoriu cele două eșantioane de lucru, include valoarea zero, astfel încât nu putem concluziona faptul că terapia de substituție nicotinică este semnificativ mai eficientă comparativ cu consilierea psihologică. Valoarea zero din interiorul intervalului de încredere indică faptul că există posibilitatea ca între cele două grupuri să nu existe nicio diferență ($\pi_1 - \pi_2 = 0 \Rightarrow \pi_1 = \pi_2$).

În concluzie, deși terapia de substituție nicotinică pare să aibă o rată de succes mai mare (28% vs. 24%), diferența evidențiată în mod descriptiv pentru loturile studiate nu este semnificativă statistic. Cu alte cuvinte, diferența de 4%, evidențiată la nivel de eșantion, nu este suficient de mare încât să putem exclude posibilitatea absenței diferenței dintre cele două proporții la nivel populațional. Astfel, poate fi luată în considerare extinderea cercetării prin utilizarea în cadrul studiului a unor eșantioane de volum superior sau o analiză statistică suplimentară, care să poată scoate în evidență potențiali factori de confuzie care ar putea denatura rezultatele (de exemplu, variabilele de control vârsta, sexul, alte tratamente medicamentoase, stresul, etc.).

Scanned with OKEN Scanner

Capitolul. 8

8. TESTAREA IPOTEZELOR STATISTICE. VOLUMUL OPTIM AL EȘANTIONULUI

SUBIECTE:

8.1. Definiția și principalele caracteristici ale unui test statistic
8.2. Erorile specifice procesului de testare a ipotezelor statistice
8.3. Determinarea volumului eșantionului de studiu pentru estimarea mediei unei populații
8.4. Determinarea volumului eșantionului de studiu pentru estimarea proporției unei populații

La finalizarea acestui capitol, studentul va fi capabil să:
O8-1: să înțeleagă cum să formuleze corect ipotezele nulă și alternativă în cadrul unui test de verificare a ipotezelor statistice structurat
O8-2: să înțeleagă conceptele de eroare de tip I și eroare de tip II
O8-3: să calculeze volumul eșantionului de studiu pentru estimarea mediei unei populații
O8-4: să calculeze volumul eșantionului de studiu pentru estimarea proporției unei populații

8.1. DEFINIȚIA ȘI PRINCIPALELE CARACTERISTICI ALE UNUI TEST STATISTIC

Similar cu situația estimării prin interval de încredere a valorilor parametrilor unei populații (media, proporția, varianța, etc.) prezentată în capitolul anterior și în cazul testării ipotezelor statistice se urmărește obținerea unei concluzii referitoare la o anumită populație supusă studiului prin examinarea unui eșantion reprezentativ extras din aceasta.

În procesul complex de luare a deciziilor din cadrul oricărui domeniu de activitate, personalul implicat trebuie să gestioneze cât mai eficient incertitudinea, neputând fi siguri niciodată de faptul că decizia pe care o luăm este „absolut" corectă.

Cu toate acestea, se poate dezvolta un set de reguli care să ne poată ajuta să controlăm și astfel să minimizăm probabilitatea de a comite o eroare în alegerea soluției optime din mai multe alternative posibile.

Scanned with OKEN Scanner

Într-o manieră simplificată, putem privi o ipoteză drept o afirmație privitoare la parametrii uneia sau mai multor populații care prezintă interes din diferite perspective. De exemplu:

- managerul unui spital poate formula o ipoteză potrivit căreia durata medie a internării pacienților dintr-un spital este de 4 zile;
- un medic specialist poate formula o ipoteză despre faptul că, un anumit medicament poate fi eficient pentru 90% dintre pacienți;
- o companie petrolieră afirmă faptul că aditivul pe care îl utilizează este asociat cu o scădere a consumului de combustibil cu 10%;
- un profesor le comunică studenților săi faptul că participarea la peste trei sferturi din activitățile de curs pe parcursul semestrului conduce la o creștere a notei finale la examen cu 15%.

Pentru o înțelegere mai bună a fenomenului, ne putem gândi la studiul eficienței unui medicament nou prin compararea rezultatelor obținute în grupurile de control și de tratament. Astfel, procedura de testare a ipotezelor statistice ne poate furniza informația potrivit căreia efectul medicamentului observat la nivelul eșantionului studiat este probabil să se regăsească și la scara întregii populații din care acesta a fost extras. Acest aspect este foarte important pentru că nu ne dorim să utilizăm pe scară largă medicamentul, dacă el este eficient doar raportat la pacienții din lotul analizat, ci avem nevoie de dovezi că va fi eficient pentru întreaga populație de referință. Procedura de testare a ipotezelor statistice ne oferă posibilitatea să putem formula aceste concluzii despre populațiile statistice în ansamblul lor.

Ipotezele statistice reprezintă o serie de presupuneri despre parametrii unor populații, care sunt formulate în așa fel încât să permită o evaluare corespunzătoare utilizând metode statistice adecvate.

Primul pas în procesul testării ipotezelor statistice este reprezentat de formularea ipotezei nule despre parametrul / parametrii unei / unor populații (H_0). *Ipoteza nulă* este cunoscută drept *ipoteza absenței diferențelor*, fiind utilizată în scopul expres de a fi discreditată. În consecință, varianta diametral opusă a ceea ce dorește să dovedească echipa de cercetare va fi folosită ca enunț al ipotezei nule.

Ne putem gândi la ipoteza nulă ca reprezentând teoria clasică (implicită), care necesită dovezi suficient de solide, furnizate de eșantionul supus studiului nostru, pentru a putea fi în consecință respinsă. De exemplu, în situația când comparăm mediile a două populații, ipoteza nulă afirmă faptul că diferența dintre cele 2 medii este egală cu zero, sau, cu alte cuvinte, populațiile supuse analizei nu sunt diferite.

În cadrul procesului de testare, ipoteza nulă poate fi respinsă sau nu. În situația când H_0 nu este respinsă vom putea afirma că *„datele care stau la baza testului nostru nu aduc suficiente dovezi în sprijinul respingerii ipotezei nule"*. Atunci când analiza furnizează suficiente probe care conduc la respingerea

Scanned with OKEN Scanner

ipotezei nule vom concluziona că datele nu sunt compatibile cu ipoteza nulă dar susțin alte ipoteze.

Ipoteza alternativă (H_1) reprezintă o afirmație despre care vom crede faptul că este adevărată în situația când analiza datelor la nivel de eșantion ne va determina să respingem ipoteza nulă. De cele mai multe ori ipoteza alternativă coincide cu ipoteza cercetării, presupunerea de bază care motivează și susține procesul creator în majoritatea studiilor științifice.

Trebuie precizat faptul că *procesul de testare a ipotezelor statistice în particular și chiar inferența statistică în general nu conduc către dovedirea vreunei ipoteze*, ci doar indică situația în care o anumită ipoteză statistică este susținută sau nu de datele disponibile la nivel de eșantion. Atunci când eșuăm în a respinge ipoteza nulă nu afirmăm că aceasta este adevărată ci că „ar putea fi adevărată".

Testul statistic reprezintă *un indicator care poate fi calculat pe baza datelor din eșantion*. Există un număr enorm de valori pe care un test statistic le poate avea, fiecare dintre acestea depinzând de eșantionul particular care a fost extras. În funcție de nivelul concret pe care îl înregistrează, prin comparație cu valoarea tabelată corespunzătoare tipului de repartiție, testul statistic servește ca *element de bază în luarea deciziei de a respinge sau nu ipoteza nulă*.

Formula generală a unui test statistic este:

$$Testul\ statistic = \frac{Statistica\ de\ esantion - Parametrul\ populatiei}{Eroarea\ standard\ a\ statisticii\ de\ esantion} \tag{8.1}$$

De exemplu, pentru media unei anumite populații, în situația în care eșantionul a fost extras dintr-o populație normal distribuită, ne putem situa într-una dintre situațiile:

a) varianța (σ^2) a populației este *cunoscută*: testul statistic utilizat pentru testarea ipotezei nule H_0: $\mu = \mu_0$ va fi:

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{8.2}$$

care, atunci când H_0 este adevărată, urmează o distribuție normală standard z;

b) varianța (σ^2) a populației este necunoscută: testul statistic utilizat pentru testarea ipotezei nule H_0: $\mu = \mu_0$ va fi:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{8.3}$$

Scanned with OKEN Scanner

care, atunci când H_0 este adevărată, urmează o distribuție Student cu $n-1$ grade de libertate.

În fiecare caz în parte (t, z, χ^2, etc.), *distribuțiile statistice de sondaj aferente presupun faptul că ipoteza nulă este corectă* pentru populația din care am extras eșantioanele în mod aleatoriu. În scopul de a evalua cât de compatibile sunt datele obținute la nivelul de eșantionului nostru cu ipoteza nulă H_0, trebuie să poziționăm valoarea calculată a testului statistic în distribuția sa de sondaj și să determinăm cât de neobișnuită este aceasta. După cum am prezentat în capitolele anterioare, distribuțiile t, z, χ^2sunt distribuții probabilistice pe care le putem utiliza pentru a calcula diferite probabilități pentru regiunile situate între curbele respective și axa absciselor.

Presupunem ca avem o statistică (media) calculată la nivelul unui eșantion ($n = 21$) care urmează o distribuție probabilistică t *(Student)* cu $n - 1 = 20$ grade de libertate. Distribuția este centrată în jurul valorii 0, deoarece presupune că ipoteza nulă H_0 a absenței diferențelor este adevărată. În situația când H_0 este adevărată, studiul nostru va genera cel mai probabil o valoare a statisticii t, calculată la nivelul eșantionului, apropiată de zero, șansele de a obține valori t mai îndepărtate de zero (în ambele sensuri) fiind din ce în ce mai reduse, pe măsură ce ne apropiem de valorile extreme.

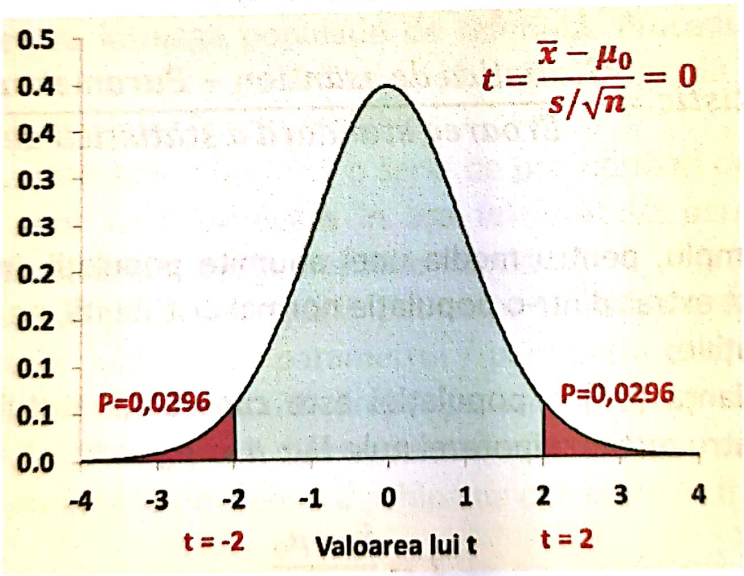


Figura 8.1. Distribuția Student „t" cu 20 de grade de libertate

Considerăm că am obținut la nivelul eșantionului o valoare calculată $t = 2$ în urma studiului ipotetic și dorim să vedem în ce măsură datele din eșantion se „suprapun" peste ipoteza nulă. Din analiza graficului prezentat în figura 8.1, observăm faptul că, considerând că ipoteza nulă este adevărată, distribuția *Student* ne indică că $t = 2$ nu este una dintre cele mai probabile valori pe care le-am putea obține (zonele marcate cu roșu, testul fiind bilateral).

Scanned with OKEN Scanner

Rezultă faptul că, valoarea $t = 2$ se obține destul de rar atunci când ipoteza nulă este adevărată. Ne interesează însă cât de rară este această valoare, scopul nostru final fiind reprezentat de *„evaluarea nivelului de raritate"* al valorii t, calculată pentru eșantionul nostru, în scopul de a putea în final justifica respingerea ipotezei nule pentru întreaga populație din care acesta a fost extras. Astfel, avem nevoie să cuantificăm probabilitatea observării valorii calculate $t = 2$.

În acest exemplu, avem un test *„t"* bilateral care ne permite să evaluăm situațiile în care media eșantionului este mai mare sau mai mică decât valoarea de referință. Graficul distribuției Student (figura 8.1) indică faptul că ambele regiuni $(-\infty; -2)$ și, respectiv, $(2; +\infty)$ evidențiază fiecare o probabilitate de 0,0296, dintr-un total de 0,0592. Astfel, valorile testului t se încadrează în aceste regiuni aproximativ 6% din timp atunci când ipoteza nulă este adevărată. Cu alte cuvinte, în situația în care am extrage în mod aleatoriu 100 de eșantioane de volum $n = 21$ din populația studiată, doar în aproximativ 6 dintre acestea am înregistra valori ale testului t mai mari sau egale cu 2 sau mai mici sau egale cu -2, iar în 94 dintre acestea, valori cuprinse între -2 și 2.

În situația testelor bilaterale (H_0: efectul este egal cu zero și H_1: efectul nu este egal cu zero) se pot detecta atât efectele negative cât și cele pozitive, această categorie de teste reprezentând un standard în procesul de cercetare științifică, unde descoperirea oricărui tip de efect este de obicei de interes pentru echipa de cercetare.

Dimensiunea efectului reprezintă *diferența dintre valoarea parametrului populației și cea din ipoteza nulă H_0 sau dintre valorile parametrilor a două sau mai multe populații.*

Pentru a exemplifica, considerăm diferența medie dintre rezultatul obținut în urma utilizării unui medicament în grupul de tratament și în grupul de control. Dacă presupunem faptul că media unui grup este 15 și media celuilalt grup este 7, atunci efectul brut este $15 - 7 = 8$. În practică, întâlnim și forma standardizată a efectului pentru diferența mediilor a două populații (Cohen's d), care se obține prin raportarea diferenței prezentate anterior la deviația standard comună, de obicei estimată în baza unor studii pilot sau preluată din rezultatele altor cercetări similare realizate anterior. Din punctul de vedere al medicului specialist, dimensiunea defectului reprezintă acea diferență clinică relevantă, care poate determina modificarea sistemului de tratament existent.

Valoarea critică a testului: reprezintă acele valori care separă regiunea/regiunile de respingere de cea de „acceptare" și ne arată cât de „extremă" trebuie să fie valoarea testului statistic determinat la nivel de eșantion, atunci când ipoteza nulă este adevărată, pentru ca probabilitatea de a obține o valoare similară sau mai mare să fie α.

De exemplu, în situația unui test z bilateral, când dorim ca probabilitatea de a respinge ipoteza nulă adevărată să fie 5% ($\alpha = 0{,}05$), o parte a lui α va fi

Scanned with OKEN Scanner

asociată cu valorile foarte mici și o parte cu valorile foarte mari (figura 8.2). În consecință, vom împărți valoarea nivelului de semnificație α în două părți egale, fiecare fiind $\alpha/2 = 0{,}025$. Astfel, în cazul distribuției normale, dorim să determinăm acel nivel al lui z la dreapta căruia se află $0{,}025$ sau 2,5% din aria situată sub graficul distribuției și la stânga căruia se situează $1 - 0{,}025 = 0{,}975$ sau 97,5% din suprafața delimitată de curbă, axa absciselor și $-\infty$ (figura 8.2).

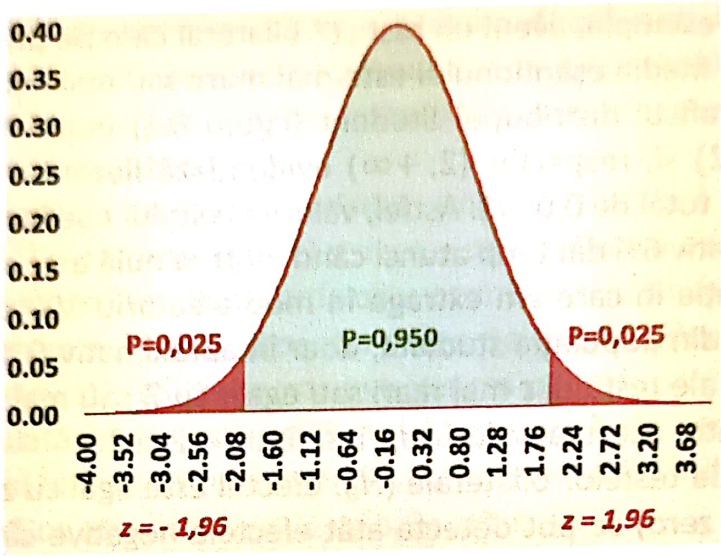


Figura 8.2. Distribuția normală (test z bilateral)

Din tabelele cu distribuția normală standard citim valoarea $z = 1{,}96$ corespunzătoare (probabilitatea obținerii unei asemenea valori sau a uneia mai mari este egală cu 0,025). Așa cum am văzut în capitolul anterior, putem utiliza funcția Excel: $NORM.S.INV(0{,}975)$. Similar, deoarece curba normală este perfect simetrică, obținem $z = -1{,}96$ (probabilitatea obținerii unei asemenea valori sau a uneia mai mici este egală cu 0,025).

Regula de decizie: toate valorile posibile pe care le poate înregistra un test statistic sunt puncte situate pe axa orizontală a graficului distribuției de sondaj a acestuia, care sunt împărțite în două grupuri:

- *regiunea de respingere*: cuprinde valorile testului statistic (z, t, F, etc.) care sunt cel mai puțin probabil să apară, atunci când ipoteza nulă (H_0) este adevărată (zonele marcate cu roșu în figura 8.2);
- *regiunea de „acceptare"*: cuprinde valorile testului statistic care apar cel mai probabil în situația când ipoteza nulă (H_0) este adevărată (aria colorată în verde din figura 8.2).

Continuând pe linia exemplului anterior, regiunea noastră de respingere cuprinde toate valorile testului statistic (z) mai mari sau egale cu 1,96 și mai mici sau egale cu -1,96. Regiunea de „acceptare" va cuprinde toate valorile cuprinse între -1,96 și 1,96. Astfel, regula de decizie ne indică să respingem ipoteza nulă (H_0) în cazul în care valoarea testului statistic calculată pe baza datelor din

Scanned with OKEN Scanner

eșantion aparține regiunii de respingere și să nu respingem ipoteza nulă dacă valoarea testului se situează în regiunea de „acceptare".

Nivelul de semnificație (α): ne indică cât de relevante trebuie să fie probele evidențiate la nivel de eșantion pentru a putea concluziona faptul că efectul determinat pentru acesta există și la nivelul întregii populații din care el a fost extras în mod aleatoriu. El specifică în ce măsură probele obținute la nivelul eșantionului trebuie să contrazică ipoteza nulă pentru a putea respinge H_0 pentru întreaga populație. Acest standard reprezintă *probabilitatea respingerii unei ipoteze nule adevărate* sau probabilitatea de a afirma că un efect există (o diferență, o corelație, o asociere) când de fapt acesta nu este prezent în populație. De exemplu, un nivel de semnificație $\alpha = 0{,}05$ semnifică un risc de 5% de a lua decizia că un anumit efect există la nivel populațional când în realitate acesta este inexistent, fiind observat din pură întâmplare la nivelul eșantionului analizat. Deoarece *respingerea unei ipoteze nule adevărate reprezintă o eroare* este necesar minimizăm probabilitatea comiterii acesteia. În consecință, valorile cel mai des întâlnite în practică pentru α sunt 0,01; 0,05 și 0,10.

Valoarea p: indică nivelul de relevanță al probelor analizate la nivel de eșantion împotriva ipotezei nule H_0, cu alte cuvinte în ce măsură datele din eșantion au puterea de a contrazice ipoteza nulă. Valoarea lui p reprezintă probabilitatea de a obține la scara întregii populații efectul observat la nivelul eșantionului sau un efect mai important, în situația când ipoteza nulă H_0 (care afirmă faptul că nu există nici un efect) este adevărată. Astfel, valoarea p reprezintă *valoarea minimă a nivelului de semnificație α pentru care putem respinge ipoteza nulă*.

În situația în care $p < \alpha$, putem respinge ipoteza nulă H_0 și în consecință afirmăm faptul că *rezultatele studiului nostru sunt „semnificative statistic"*. Altfel spus, datele din eșantionul analizat susțin ipoteza alternativă H_1, potrivit căreia efectul observat este prezent și în cadrul populației generale. Atunci când valoarea lui p este superioară nivelului de semnificație α *($p > \alpha$)*, datele obținute la nivel de eșantion nu aduc suficiente dovezi pentru a concluziona faptul că efectul există în populația studiată. Deși la prima vedere afirmația *„nu am reușit să respingem ipoteza nulă"* implică o dublă negație trebuie să evităm formularea acesteia sub forma „acceptăm ipoteza nulă", deoarece nu putem dovedi o negație, probele insuficiente nefiind o dovadă a faptului că ceva nu există. *Nu am reușit să dovedim că efectul există, acesta ar putea exista, dar studiul nostru nu a reușit să-l evidențieze*.

Este foarte utilă *analogia cu procesele penale*, care au drept punct de plecare ipoteza că acuzatul este nevinovat, până i se dovedește vinovăția. Procurorul trebuie să depună un efort însemnat în scopul de a depăși un standard de evidență necesar obținerii unui verdict de vinovăție. În situația în care procurorul nu reușește, acest lucru nu reprezintă o dovadă a faptului că acuzatul este inocent, ci doar că au fost aduse insuficiente probe în sprijinul vinovăției sale.

Scanned with OKEN Scanner

Este posibil, fie ca investigația condusă de procuror să fi avut un caracter superficial și în consecință să fi fost pierdute probe, fie ca acuzatul să-și fi acoperit cu succes urmele (Navarro Danielle, Foxcroft David, Learning Statistics with Jamovi). Astfel, judecătorul nu spune că acuzatul a fost dovedit nevinovat („acceptăm ipoteza nulă H_0") ci doar că nu au fost suficiente probe care să înlăture prezumția de nevinovăție și să conducă la condamnarea acestuia (nu am reușit să respingem ipoteza nulă H_0).

8.2. ERORILE SPECIFICE PROCESULUI DE TESTARE A IPOTEZELOR STATISTICE

Așa cum am văzut anterior, datele analizate la nivel de eșantion trebuie să aducă suficiente dovezi pentru a putea respinge ipoteza nulă H_0 și, în consecință, pentru a putea concluziona că efectul observat în eșantion există și la nivelul populației din care acesta a fost extras.

Într-o situație ideală, un test de ipoteză nu conduce la respingerea ipotezei nule atunci când efectul nu este prezent la nivel populațional și respinge H_0 în situația când efectul există.

În realitate însă, în procesul de testare a ipotezelor, statisticienii se confruntă cu două tipuri de erori, numite *erori de tipul I* și *erori de tipul II*, ambele fiind relaționate cu concluzii incorecte privitoare la ipoteza nulă H_0 (tabelul 8.1).

Tabelul 8.1. Tipuri de erori specifice procesului de testare a ipotezelor statistice

Rezultatul testului / Situație de fapt	Respingem Ipoteza nulă H_0	Nu respingem Ipoteza nulă H_0
Ipoteza nulă H_0 este adevărată	Eroare de tip I (α) (fals pozitiv)	Decizie corectă ($1-\alpha$) (nu avem efect)
Ipoteza nulă H_0 este falsă	Decizie corectă ($1-\beta$) (efectul există)	Eroare de tip II (β) (fals negativ)

Erorile de tipul I (valorile fals pozitive): deși atunci când obținem o valoare p inferioară nivelului de semnificație fixat α ($p \leq \alpha$) considerăm că rezultatele studiului nostru sunt semnificative statistic, putem fi în situația comiterii unei erori de tip I, adică efectul observat la nivel de eșantion nu se regăsește în cadrul populației generale. Atunci când ne situăm în această ipostază *nu suntem în mod obișnuit avertizați în nici un fel*. Erorile de tip I se datorează erorilor de sondaj, eșantionul nostru supraestimând efectul studiat din pură întâmplare. Chiar dacă

Scanned with OKEN Scanner

nu putem distinge care dintre studii au rezultate fals pozitive, cunoaștem în schimb rata de apariție a erorilor de tip I, care este egală cu nivelul de semnificație al testului statistic (α). Valoarea setată pentru α ține de cât de mult suntem dispuși să riscăm apariția unei valori fals pozitive. Astfel, în situația când stabilim nivelul de semnificație $\alpha = 0{,}05$, iar ipoteza nulă H_0 este adevărată, există o șansă de 5% ca testul nostru să respingă H_0 în mod incorect sau să comitem o eroare de tip I (*rata fals pozitivilor este de 5%*).

Erorile de tip II (valorile fals negative): în situația în care aplicăm un test de ipoteză și valoarea p este mai mare decât nivelul de semnificație α ($p > \alpha$), rezultatele noastre nu sunt semnificative statistic, eșantionul analizat nefurnizând suficiente dovezi că efectul studiat există în populația generală. Cu toate acestea, există o șansă ca efectul să fie prezent în cadrul populației, cu toate că rezultatele testului statistic aplicat nu susțin această ipoteză. Această situație apare în practică atunci când ne confruntăm cu o eroare de tip II (β).

În contrast cu situația erorilor de tip I, a care apar doar ca urmare a erorilor de sondaj, în situația erorilor de tip II avem mai mulți factori cauzali (eșantioane de volum mic, efecte de dimensiuni reduse, un grad mare de variabilitate în populația generală). În plus, în mod diferit față de erorile de tip I, în cazul erorilor de tip II nu există posibilitatea stabilirii de la început a unui nivel standard, ci doar estimarea valorii acestora anterior începerii studiului prin aproximarea proprietăților ipotezei alternative (H_1).

Rata erorilor de tip II (β) reprezintă probabilitatea obținerii unui rezultat fals negativ al testului statistic. În mod firesc, inversul erorilor de tip II reprezintă probabilitatea detectării unui efect în mod corect sau *puterea testului statistic* ($1 - \beta$).

Puterea unui test statistic este influențată în mod direct de trei factori: dimensiunea eșantionului, nivelul variabilității existent în populație și dimensiunea efectului analizat.

Trebuie să ținem cont de faptul că nivelul variabilității (evidențiat prin deviația standard a populației) și dimensiunea efectului existent în colectivitatea generală reprezintă doar aproximări și presupuneri și, în consecință, puterea testului și rata erorilor de tip II sunt și ele doar valori estimative și nu ceva ce putem alege în mod direct, ca în cazul nivelului de semnificație α.

Un nivel redus al variabilității datelor din populația generală sau un efect de o dimensiune mai mare determină o diminuare a erorilor de tip II și, în consecință, o putere mai mare a testului statistic. Din nefericire, echipa de cercetare are un control mai redus asupra acestor aspecte, astfel încât *elementul cel mai important prin care putem controla efectul erorilor de tip II* este reprezentat de *dimensiunea eșantionului* care va fi utilizat în cadrul studiului. Păstrând constante toate celelalte elemente, prin creșterea volumului eșantionului de lucru reducem rata de apariție a erorilor de tip II și implicit majorăm puterea testului utilizat.

Scanned with OKEN Scanner

Graficul 8.3 ilustrează cele două tipuri de erori utilizând două distribuții de sondaj, una pentru ipoteza nulă H_0 adevărată (în stânga) și cealaltă pentru ipoteza alternativă H_1 adevărată (în dreapta). Linia verticală care indică nivelul critic sau pragul decizional reprezintă punctul în respingem sau eșuăm în a respinge ipoteza nulă. Atunci când ipoteza nulă este adevărată trebuie să avem în vedere doar erorile de tip I, respectiv zona marcată cu roșu din cadrul figurii. De asemenea, în situația când ipoteza alternativă este adevărată, atenția trebuie concentrată asupra erorilor de tip II (zona marcată cu albastru).

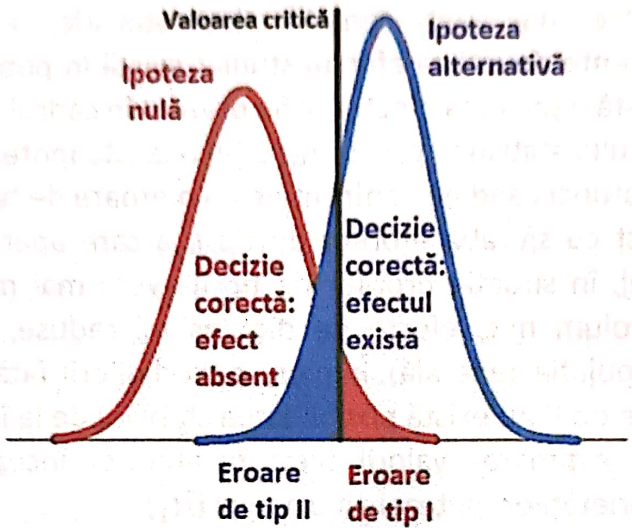


Figura 8.3. Erorile de tip I și II

Deplasarea liniei valorii critice este echivalentă cu modificarea nivelului de semnificație α. Astfel, deplasând linia spre stânga realizăm o creștere a nivelului de semnificație (de ex. α se modifică de la 0,05 la 0,10) și astfel rata erorii de tip I devine mai mare, în timp ce rata erorii de tip II (β) se reduce în mod inevitabil. Mutând în schimb linia spre dreapta, micșorăm nivelul de semnificație (α se modifică de la 0,05 la 0,01), determinând o reducere a ratei erorilor de tip I, realizată însă pe fondul creșterii ratei erorilor de tip II.

Se observă în mod evident dependența invers proporțională între cele 2 tipuri de erori, care reprezintă de fapt probabilități. Pentru a micșora ambele erori este necesară *folosirea unor eșantioane de volum mare*, care vor determina în special reducerea erorii de tip II. Cu cât cele 2 distribuții se suprapun mai puțin (dimensiunea efectului este mai mare), cu atât erorile de decizie sunt mai reduse.

Scanned with OKEN Scanner

8.3. DETERMINAREA VOLUMULUI EȘANTIONULUI DE STUDIU PENTRU ESTIMAREA MEDIEI UNEI POPULAȚII

În fiecare situație în parte, atunci când lucrăm cu eșantioane și nu cu populațiile generale din care acestea provin, este foarte important să avem siguranța faptului că eșantionul are o dimensiune optimă și reprezintă în mod adecvat întreaga populație. Dacă nu ținem cont de aceste două aspecte fundamentale, rezultatele obținute în urma prelucrării datelor la nivel de eșantion nu pot fi generalizate pentru populația care face obiectul studiului nostru.

Scopul principal al calculării volumului optim al eșantionului în cazul unui studiu clinic este reprezentat de *determinarea numărului necesar de participanți pentru a fi în măsură să detectăm un efect* (o diferență, o asociere, o corelație) *relevant în cazul unui tratament clinic*. Utilizarea unui eșantion de volum superior celui necesar pentru a obține rezultatele dorite reprezintă o modalitate de irosire în mod inutil a resurselor, în timp ce, la popul opus, eșantioanele de volum prea redus tind deseori să genereze rezultate fără o utilitate practică.

Vom prezenta în continuare o modalitate simplă de determinare a volumului eșantionului de lucru necesar pentru estimarea mediei unei populații, în cazul unei variabile continue. După cum am detaliat în capitolul anterior, privitor la estimarea parametrilor unei populații și intervalele de încredere, obiectivul principal în cadrul procesului de estimare este reprezentat de obținerea unor intervale de variație cât mai înguste în care să avem încredere (de obicei $1-\alpha=95\%$) că se află valoarea adevărată dar necunoscută a parametrului de interes, în cazul nostru media populației generale.

Conform teoremei limită centrală, în situația când volumul eșantionului este mare ($n \geq 30$) distribuția de sondaj a mediilor eșantioanelor poate fi aproximată printr-o distribuție normală având media μ și deviația standard $\sigma/\sqrt{n}$. În baza relației 7.10, prezentată în capitolul anterior, valoarea $z_{\bar{x}}$ corespunzătoare mediei eșantionului $\bar{x}$ extras din populația generală cu media μ este:

$$z_{\bar{x}} = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

Valoarea lui $z_{\bar{x}}$ este situată în intervalul simetric delimitat de valorile critice $z_{\alpha_1} = -z_{1-\frac{\alpha}{2}}$ și $z_{\alpha_2} = +z_{1-\frac{\alpha}{2}}$ (relația 7.11):

$$-z_{1-\frac{\alpha}{2}} \leq z_{\bar{x}} \leq z_{1-\frac{\alpha}{2}}$$

Înlocuind valoarea lui $z_{\bar{x}}$, obținem:

Scanned with OKEN Scanner

$$-z_{1-\frac{\alpha}{2}} \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{1-\frac{\alpha}{2}} \Rightarrow -z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{x}-\mu \leq z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Volumul eșantionului se determină punând condiția ca jumătate din dimensiunea intervalului de confidență pentru media populației generale ($z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$) să fie mai mică decât eroarea maximă admisă de specialiști (d).

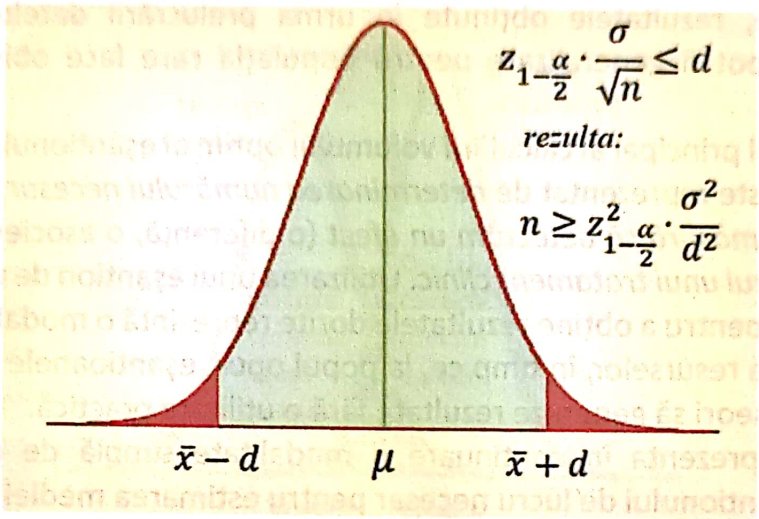


Figura 8.4. Determinarea volumului eșantionului pentru estimarea mediei unei populații

Eroarea maximă admisă (d) în cadrul fiecărui studiu este stabilită anterior de către echipa de cercetare, aceasta reprezentând diferența maximă acceptată între valoarea parametrului populației generale (μ) și valoarea estimatorului său punctual, calculat la nivel de eșantion ($\bar{x}$).

Astfel, avem:

$$\mu - \bar{x} \leq d \quad \Rightarrow \quad z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq d$$

Rezultă imediat relația de calcul a volumului eșantionului în situația estimării mediei unei populații:

$$n \geq z^2_{1-\frac{\alpha}{2}} \cdot \frac{\sigma^2}{d^2} \qquad (8.1)$$

Vom prezenta în continuare un prim exemplu privind estimarea numărului mediu de dinți afectați de carii pentru copiii având vârstele cuprinse între 6 și 8 ani. O echipă de medici stomatologi dorește să estimeze numărul

Scanned with OKEN Scanner

mediu de dinți afectați de carii la copiii cu vârstele cuprinse între 6 și 8 ani, cu o precizie de $\pm 0{,}5$ dinți, utilizând un nivel de încredere $1 - \alpha = 95\%$.

Cunoaștem din studii realizate anterior deviația standard estimată a numărului de dinți afectați de carii din cadrul populației generale, respectiv $\sigma = 1{,}5$ dinți. Eroarea maximă admisibilă, stabilită de echipa de medici, este $d = 0{,}5$ dinți, iar nivelul de încredere pentru estimarea numărului mediu de dinți afectați de carii folosit în cadrul acestei cercetări este $1 - \alpha = 95\%$. Valoarea critică, citită din tabelul distribuției normale standard, este $z_{1-\frac{\alpha}{2}} = 1{,}96$. Înlocuind aceste date în relația 8.1, obținem:

$$n \geq z^2_{1-\frac{\alpha}{2}} \cdot \frac{\sigma^2}{d^2} \Rightarrow n \geq 1{,}96^2 \cdot \frac{1{,}5^2}{0{,}5^2} \Rightarrow n \geq 34{,}57$$

Rotunjind în plus, vom avea nevoie de un eșantion compus din cel puțin 35 de copii, cu vârstele cuprinse între 6 și 8 ani, pentru a estima media numărului de dinți afectați de carii cu eroare de $\pm 0{,}5$ și un nivel de încredere de 95%. Un astfel de calcul este foarte important în situațiile când se planifică diferite studii epidemiologice în școli sau când se elaborează o serie de programe de prevenție dentară pentru copiii înscriși în învățământul primar.

În continuare (tabelul 8.2) vom determina variația volumului eșantionului pe măsura creșterii valorii erorii limită d, în cadrul unui interval cuprins între 0,1 și 1 dinți ($\alpha = 5\%$ și $\sigma = 1{,}5$ dinți rămân nemodificate). Astfel:

Tabelul 8.2. Variația volumului eșantionului pe măsura creșterii valorii erorii limită d în intervalul cuprins între 0.1 și 1 dinți

Eroarea maximă admisă d	Nivelul de încredere $1 - \alpha$	Valoarea critică $z_{1-\frac{\alpha}{2}}$	Deviația standard estimată a populației generale σ	Volumul minim al eșantionului $n \geq z^2_{1-\frac{\alpha}{2}} \cdot \frac{\sigma^2}{d^2}$
0,1	95%	1,96	1,5	865
0,2	95%	1,96	1,5	217
0,3	95%	1,96	1,5	97
0,4	95%	1,96	1,5	55
0,5	95%	1,96	1,5	35
0,6	95%	1,96	1,5	25
0,7	95%	1,96	1,5	18
0,8	95%	1,96	1,5	14
0,9	95%	1,96	1,5	11
1,0	95%	1,96	1,5	9

Scanned with OKEN Scanner

Pe măsura creșterii erorii maxime admise d, de la valoarea 0,1 la valoarea 1, volumul eșantionului se diminuează „dramatic" de la 865 copii, cu vârstele cuprinse între 6 și 8 ani, la doar 9 copii din acest interval de vârstă. Cu alte cuvinte, pe măsură ce suntem dispuși să acceptăm (tolerăm) un nivel mai ridicat al erorii maxime admise d, volumul minim al eșantionului nostru n, necesar în vederea estimării numărului mediu de dinți afectați de carii pentru copiii având vârstele cuprinse între 6 și 8 ani, se diminuează considerabil.

În al doilea rând, vom evidenția variația volumului eșantionului pe măsura creșterii nivelului de încredere $(1-\alpha)$, în cadrul unui interval valoric cuprins între 90 și 99 (tabelul 8.3), restul indicatorilor rămânând nemodificați (eroarea maximă admisă $d = 0{,}5$ dinți și $\sigma = 1{,}5$ dinți).

Tabelul 8.3. Variația volumului eșantionului pe măsura creșterii nivelului de încredere $1-\alpha$ de la 90 până la 99

Eroarea maximă admisă d	Nivelul de încredere $1-\alpha$	Valoarea critică $z_{1-\frac{\alpha}{2}}$	Deviația standard estimată a populației generale σ	Volumul minim al eșantionului $n \geq z^2_{1-\frac{\alpha}{2}} \cdot \frac{\sigma^2}{d^2}$
0,5	90%	1,64	1,5	24,35
0,5	91%	1,70	1,5	25,87
0,5	92%	1,75	1,5	27,58
0,5	93%	1,81	1,5	29,55
0,5	94%	1,88	1,5	31,84
0,5	95%	1,96	1,5	34,57
0,5	96%	2,05	1,5	37,96
0,5	97%	2,17	1,5	42,38
0,5	98%	2,33	1,5	48,71
0,5	99%	2,58	1,5	59,71

Din analiza tabelului 8.3, rezultă imediat faptul că, pe măsura creșterii nivelului de încredere de la 90% până la 99%, păstrând constante eroarea maximă admisă la nivelul $d = 0{,}5$ dinți și deviația standard estimată a populației generale $\sigma = 1{,}5$ dinți, volumul minim al eșantionului se majorează de la 25 de copii cu vârstele cuprinse între 6 și 8 ani, pentru un nivel de încredere de 90%, până la 60 de copii, în cazul unui nivel de încredere de 99%.

În al treilea rând, vom evalua variația volumului eșantionului pe măsura creșterii deviației standard a populației de la valoarea de 1,5 până la 2,5 dinți (tabelul 8.4), restul indicatorilor rămânând nemodificați ($\alpha = 5\%$ și $d = 0{,}5$ dinți).

Scanned with OKEN Scanner

Tabelul 8.4. Variația volumului eșantionului pe măsura creșterii deviației standard a populației σ de la 1,5 până la 2,5 dinți

Eroarea maximă admisă d	Nivelul de încredere $1-\alpha$	Valoarea critică $z_{1-\frac{\alpha}{2}}$	Deviația standard estimată a populației generale σ	Volumul minim al eșantionului $n \geq z^2_{1-\frac{\alpha}{2}} \cdot \frac{\sigma^2}{d^2}$
0,5	95%	1,96	1,5	34,57
0,5	95%	1,96	1,6	39,34
0,5	95%	1,96	1,7	44,41
0,5	95%	1,96	1,8	49,79
0,5	95%	1,96	1,9	55,47
0,5	95%	1,96	2,0	61,46
0,5	95%	1,96	2,1	67,76
0,5	95%	1,96	2,2	74,37
0,5	95%	1,96	2,3	81,29
0,5	95%	1,96	2,4	88,51
0,5	95%	1,96	2,5	96,04

Conform datelor prezentate în tabelul 8.4, se observă o legătură direct proporțională între nivelul dispersiei din populația generală, cuantificat prin deviația standard, și nivelul minim necesar al eșantionului, în vederea estimării numărului mediu de dinți afectați de carii pentru copiii având vârstele cuprinse între 6 și 8 ani. Prin urmare, pentru un nivel mai redus al împrăștierii din populația generală, $\sigma = 1{,}5$ dinți, volumul minim al eșantionului trebuie să fie de 35 de copii, în timp ce, la polul opus, pentru un nivel al dispersiei la nivel populațional $\sigma = 2{,}5$ dinți, eșantionul va cuprinde cel puțin 97 de copii, cu vârstele cuprinse între 6 și 8 ani.

Considerăm un al doilea exemplu pentru determinarea volumului minim al eșantionului de studiu pentru estimarea mediei unei populații: un medic nutriționist, care dorește să realizeze un sondaj în cadrul unei populații de adolescente pentru a determina aportul alimentar mediu zilnic de proteine (măsurat în grame), cere sfatul unui biostatistician în legătura cu dimensiunea corespunzătoare a eșantionului de lucru. În acest scop, medicul nutriționist trebuie să furnizeze următoarele informații specifice: mărimea dorită a intervalului de confidență, nivelul de încredere utilizat pentru estimarea parametrului populației $(1-\alpha)$ și nivelul de variabilitate existent în populația generală (σ).

Să presupunem că medicul nutriționist ar dori un interval de aproximativ 10 grame, astfel încât estimarea să se extindă atât în stânga cât și în dreapta

Scanned with OKEN Scanner

mediei populației generale cu ± 5 grame. Cu alte cuvinte, medicul dorește o eroare maximă admisibilă $d = 5$ grame. Considerăm faptul că, în cazul acestei cercetări, se dorește un nivel de încredere $1 - \alpha = 95\%$ și de asemenea, din studiile anterioare, medicul consideră că deviația standard a populației generale σ este de aproximativ 20 de grame.

Pentru un nivel de încredere: $1 - \alpha = 95\%$ avem $z_{1-\alpha/2} = 1{,}96$; $d = 5$ grame; $\sigma = 20$ grame, volumul eșantionului va fi:

$$n \geq z^2_{1-\frac{\alpha}{2}} \cdot \frac{\sigma^2}{d^2} \Rightarrow n \geq (1{,}96)^2 \cdot \frac{20^2}{5^2} \Rightarrow n \geq 61{,}47$$

În consecință, pentru derularea studiului privind determinarea aportului alimentar mediu zilnic de proteine în cadrul populației de adolescente, medicului i se sugerează să extragă un eșantion care să cuprindă cel puțin 62 de persoane.

8.4. DETERMINAREA VOLUMULUI EȘANTIONULUI DE STUDIU PENTRU ESTIMAREA PROPORȚIEI UNEI POPULAȚII

Metoda de determinare a volumului eșantionului necesar pentru estimarea proporției unei populații este similară cu cea prezentată anterior în cazul estimării mediei.

Volumul eșantionului se determină punând condiția ca jumătate din dimensiunea intervalului de confidență pentru media populației generale ($z_{1-\frac{\alpha}{2}} \cdot \sigma_p$) să fie mai mică decât eroarea maximă admisă de specialiști (d).

Conform teoremei limită centrală, prezentate în capitolul 6, în situația când volumul eșantioanelor este suficient de mare ($n \cdot p \geq 5$ *și* $n \cdot (1-p) \geq 5$) distribuția de sondaj a tuturor proporțiilor eșantioanelor de volum n extrase dintr-o populație de volum N tinde către repartiția normală.

Dacă populația generală (N) are o proporție π și o deviație standard $\sigma = \sqrt{\pi \cdot (1-\pi)}$, atunci distribuția de sondaj a proporțiilor tuturor eșantioanelor de volum n extrase aleatoriu din aceasta are de asemenea o medie a proporțiilor π și o eroare standard a proporțiilor (deviație standard a distribuției de sondaj a proporțiilor eșantioanelor σ_p) egală cu:

$$\sigma_p = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\pi \cdot (1-\pi)}{n}}$$

Așa cum am văzut anterior, eroarea maximă admisă (d) în cadrul fiecărui studiu este stabilită anterior de către echipa de cercetare, aceasta reprezentând

Scanned with OKEN Scanner

diferența maximă acceptată între valoarea parametrului populației generale (π) și valoarea estimatorului său punctual, calculat la nivel de eșantion (p).

Astfel, avem:

$$\pi - p \leq d \quad \Rightarrow \quad z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \leq d \quad \Rightarrow \quad z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\pi \cdot (1-\pi)}{n}} \leq d$$

Rezultă imediat relația de calcul a volumului eșantionului în situația estimării proporției π a unei populații:

$$n \geq z^2_{1-\frac{\alpha}{2}} \cdot \frac{\pi \cdot (1-\pi)}{d^2} \tag{8.2}$$

Relația anterioară este utilizată în situația când populația studiată are un volum N foarte mare sau infinit. Atunci însă când populația are un volum finit, notat cu N, se aplică o corecție valorii n:

$$n_{corectat} = \frac{n}{1 + \frac{n-1}{N}} \tag{8.3}$$

Pentru determinarea practică a volumului eșantionului avem nevoie să cunoaștem proporția π din populația generală. Aceasta poate fi estimată în urma efectuării unui studiu pilot. În situația în care însă, nu putem obține un estimator pentru proporția populației, putem realiza o aproximare prin adaos, utilizând valoarea $\pi = 0{,}5$, pentru care produsul $\pi \cdot (1-\pi) = 0{,}5 \cdot 0{,}5 = 0{,}25$ înregistrează nivelul maxim (tabelul 8.5).

Tabelul 8.5. Valorile dispersiei $\sigma^2 = \pi \cdot (1-\pi)$ din populația generală

π	$\sigma^2 = \pi \cdot (1-\pi)$
0,10	0,10 * 0,90 = 0,09
0,20	0,20 * 0,80 = 0,16
0,30	0,30 * 0,70 = 0,21
0,40	0,40 * 0,60 = 0,24
0,50	**0,50 * 0,50 = 0,25**
0,60	0,60 * 0,40 = 0,24
0,70	0,70 * 0,30 = 0,21
0,80	0,80 * 0,20 = 0,16
0,90	0,90 * 0,10 = 0,09

Scanned with OKEN Scanner

Prin urmare, chiar dacă nu cunoaștem valoarea exactă a proporției π din populația generală putem determina volumul eșantionului de studiu, care va avea în acest caz o dimensiune ușor crescută. În consecință, în situația în care valoarea proporției populației va fi diferită de 0,50 (mai mică sau mai mare) precizia estimării va fi superioară celei anticipate.

Exemplul 1: managerul unui spital dorește să cunoască procentul pacienților externați care au fost nemulțumiți de calitatea serviciilor medicale de care au beneficiat în timpul perioadei de spitalizare. Ce dimensiune ar trebui să aibă eșantionul de studiu dacă dorim un nivel al erorii maxime admisibile de $\pm 5\%$ ($d = 0{,}05$), un nivel de încredere de 0,95 ($1 - \alpha = 95\%$) și nu avem nicio informație disponibilă privind proporția pacienților nemulțumiți din populația generală. Ce volum ar trebui să aibă eșantionul extras, dacă am avea conform unor cercetări anterioare informația că proporția pacienților nemulțumiți este de aproximativ $\pi = 0{,}25$?

Pentru un nivel de încredere: $1 - \alpha = 95\%$ avem $z_{1-\alpha/2} = 1{,}96$.

Deoarece nu cunoaștem valoarea proporției π existentă în cadrul populației vom estima prin adaos $\pi = 0{,}5$

Dacă eroarea de sondaj este de maxim ± 5% ($d = 0{,}05$), volumul minim al eșantionului va fi (relația 8.2):

$$n \geq z^2_{1-\frac{\alpha}{2}} \cdot \frac{\pi \cdot (1-\pi)}{d^2} \Rightarrow n \geq (1{,}96)^2 \cdot \frac{0{,}5 \cdot (1-0{,}5)}{(0{,}05)^2} \Rightarrow n \geq 384{,}16$$

În consecință, volumul minim al eșantionului necesar pentru a estima procentul pacienților externați, care au fost nemulțumiți de calitatea serviciilor medicale de care au beneficiat în timpul perioadei de spitalizare, este de 385 de pacienți.

În situația când este cunoscut volumul populației (totalitatea pacienților internați în spital), de exemplu $N = 3000$, vom aplica formula 8.3:

$$n_{corectat} = \frac{385}{1 + \frac{385 - 1}{3000}} = 341{,}31$$

În acest caz, volumul minim corectat al eșantionului va fi de 342 pacienți.

Dacă se presupune cunoscută proporția pacienților nemulțumiți ($\pi = 0{,}25$), cu o eroare de sondaj este de maxim de ± 5% ($d = 0{,}05$) , volumul minim necesar al eșantionului va fi:

$$n \geq z^2_{1-\frac{\alpha}{2}} \cdot \frac{\pi \cdot (1-\pi)}{d^2} \Rightarrow n \geq (1{,}96)^2 \cdot \frac{0{,}25 \cdot (1-0{,}25)}{(0{,}05)^2} \Rightarrow n \geq 288{,}12$$

Scanned with OKEN Scanner

Rezultă faptul că eșantionul minim necesar în vederea realizării studiului trebuie să fie compus din cel puțin 289 pacienți.

Exemplul 2: un spital desfășoară o cercetare pentru a evalua proporția pacienților dintr-un spital care dezvoltă reacții adverse în urma administrării unui nou medicament pentru hipertensiune. Managerul spitalului dorește să estimeze această proporție cu o eroare maximă de $\pm 6\%$ ($d = 0{,}06$) și un nivel de încredere $1 - \alpha = 95\%$ ($z_{1-\alpha/2} = 1{,}96$).

Dacă nu există informații anterioare despre proporția pacienților din spital cu reacții adverse la acest medicament vom estima prin adaos această valoare: $\pi = 0{,}5$

Utilizând formula 8.2, obținem:

$$n \geq (1{,}96)^2 \cdot \frac{0{,}5 \cdot (1 - 0{,}5)}{(0{,}06)^2} \Rightarrow n \geq 266{,}78$$

Prin urmare, pentru a evalua proporția pacienților dintr-un spital care dezvoltă reacții adverse la noul medicament în cadrul studiului, vom avea nevoie de un eșantion format din cel puțin 267 de pacienți.

În situația în care este cunoscută din diferite studii clinice anterioare proporția pacienților cu reacții adverse la medicament, de exemplu 10% ($p = 0{,}10$), volumul minim al eșantionului, determinat tot prin relația 8.2, va fi:

$$n \geq (1{,}96)^2 \cdot \frac{0{,}1 \cdot (1 - 0{,}1)}{(0{,}06)^2} \Rightarrow n \geq 96{,}04$$

Rotunjind în plus, obținem un volum al eșantionului de lucru de cel puțin 97 de pacienți, de aproape trei ori mai redus comparativ cu situația în care nu cunoaștem proporția existentă în populația generală.

În concluzie, atunci când există informații despre proporția reală existentă în populația analizată, obținute cu ocazia unor cercetări anterioare, putem reduce în mod semnificativ numărul de participanți din cadrul studiului, realizând în mod evident o economie importantă de timp și resurse.

Scanned with OKEN Scanner

Capitolul. 9

9. TESTELE PARAMETRICE DE VERIFICARE A IPOTEZELOR STATISTICE

SUBIECTE:

9.1. Definiția, etapele și scopul unui test parametric de verificare a ipotezelor statistice
9.2. Testarea ipotezelor pentru media unei populații - testele „z" și „t"
9.3. Testarea ipotezelor pentru proporția unei populații - testul „z"
9.4. Testarea ipotezelor pentru diferența mediilor a două populații - testele „z" și „t"
9.5. Testarea ipotezelor pentru diferența proporțiilor a două populații - testul „z"
9.6. Testarea ipotezelor pentru raportul variațelor a două populații - testul „F"

La finalizarea acestui capitol, studentul va fi capabil să:
O9-1: să înțeleagă situațiile când se utilizează în practică testele parametrice de verificare a ipotezelor statistice precum și scopul acestora
O9-2: să înțeleagă etapele unui test de verificare a ipotezelor statistice
O9-3: să calculeze și să interpreteze testele statistice „z", „t", și „F" în cadrul procesului de inferență statistică
O9-4: să înțeleagă cum să calculeze și să interpreteze valorile „p" asociate unui anumit test statistic

9.1. DEFINIȚIA, ETAPELE SI SCOPUL UNUI TEST PARAMETRIC DE VERIFICARE A IPOTEZELOR STATISTICE

Testele parametrice sunt proiectate în general pentru analiza datelor cantitative, a căror distribuție de valori respectă legea de repartiție normală sau în cazurile în care devine aplicabilă teorema limită centrală. Aceste teste de ipoteză sau de semnificație statistică compară parametrii statistici din populațiile generale studiate, cum ar fi media aritmetică, proporția sau varianța acestora.

Testele de verificare a ipotezelor statistice se realizează în baza datelor furnizate de eșantioanele de lucru, iar ulterior rezultatele obținute se extind la scara populației generale din care eșantionul a fost extras în mod aleatoriu. Această concluzie implică însă un anumit risc, stabilit în majoritatea studiilor la

Scanned with OKEN Scanner

un nivel $\alpha = 5\%$, valoare care evidențiază de fapt probabilitatea de a lua o decizie eronată, care este asumată anterior demarării cercetării. Riscul în această situație este datorat în esență imposibilității analizării tuturor unităților statistice, împreună cu caracteristicile lor, care sunt de interes în cadrul unui anumit studiu.

Testele parametrice de ipoteză pot fi bilaterale (regiunea de respingere este împărțită în două părți egale, între cele 2 cozi laterale ale distribuției) sau unilaterale (regiunea de respingere este doar într-una dintre cele 2 părți).

Toate valorile posibile pe care le poate lua un test statistic sunt puncte pe axa orizontală a graficului distribuției lui, împărțite în două grupuri: zona de respingere și zona de non-respingere (de exemplu, testul z – figura 9.1).

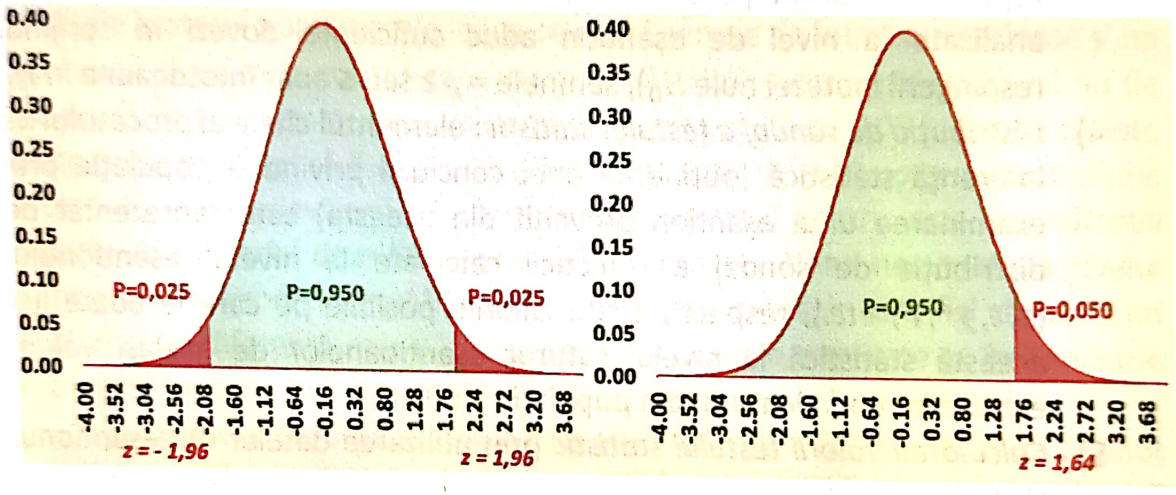


Figura 9.1. Valorile testului statistic în cazul distribuției normale standard (bilateral și unilateral dreapta)

Nivelurile testului situate în zona de respingere (marcate cu roșu în figura 9.1) reprezintă acele valori care *sunt cel mai puțin probabil să apară în cazul în care ipoteza nulă H_0 a absenței diferențelor este adevărată*.

Valorile testului care se vor situa în zona de respingere sunt calculate pe baza nivelului de semnificație α (aria de sub curba distribuției testului care depășește valorile de pe axa absciselor din zona de respingere). Cele mai frecvente valori sunt: $\alpha = 0{,}05$ și $\alpha = 0{,}01$. Nivelul de semnificație α este probabilitatea de a respinge o ipoteză nulă adevărată (*eroarea de tip I*)

Scopul testelor de verificare a ipotezelor statistice este reprezentat de asistarea factorilor de decizie (administratori, medici) în *luarea unor decizii eficiente*. Aceștia depind de decizia statistică: în situația în care ipoteza nulă este respinsă acest lucru se reflectă în faptul că decizia finală este compatibilă cu ipoteza alternativă. Și reversul este de multe ori adevărat, în situația când ipoteza nulă nu este respinsă. În această situație, este posibil însă ca factorii de decizie să considere utilă colectarea mai multor date în vederea unei analize ulterioare. Rezultatul unui test statistic reprezintă doar un singur element care stă la baza

Scanned with OKEN Scanner

unor decizii eficiente. *Decizia statistică nu trebuie interpretată ca fiind definitivă*, aceasta trebuind a fi coroborată cu celelalte informații relevante disponibile.

Principalele *etape* ale unui test parametric de verificare a ipotezelor statistice sunt:

1) *Evaluarea tipului de date* care vor fi analizate (numerice, categoriale)
2) *Presupunerile testelor* parametrice: normalitatea distribuției populațiilor din care provin eșantioanele, egalitatea variațelor populațiilor, independența eșantioanelor, etc.
3) *Ipotezele de lucru* (H_0 „ipoteza nulă sau ipoteza absenței diferențelor" va fi testată în scopul expres de a fi discreditată; H_1 „ipoteza alternativă" reprezintă o afirmație pe care o vom considera adevărată dacă datele analizate la nivel de eșantion aduc suficiente dovezi in sprijinul respingerii ipotezei nule H_0); semnele = , ≥ sau ≤ apar întotdeauna în H_0
4) *Distribuția de sondaj a testului statistic*: elementul cheie al procesului de inferență statistică (obținerea unei concluzii privind o populație prin examinarea unui eșantion provenit din aceasta) este reprezentat de distribuția de sondaj a statisticii calculate la nivelul eșantionului (z, t, χ^2, F, *etc.*), respectiv toate valorile posibile pe care le poate lua această statistică la nivelul tuturor eșantioanelor de același volum extrase în mod aleatoriu din populația studiată.
5) *Calcularea valorii testului statistic* prin utilizarea datelor din eșantionul studiat:

$$test = \frac{statistica\ din\ esantion - parametrul\ corespunzator\ al\ populatiei}{eroarea\ standard\ a\ statisticii\ din\ esantion} \quad (9.1)$$

6) *Regula de decizie*: dacă valoarea calculată a testului statistic se află în zona de respingere vom respinge ipoteza nulă, iar dacă se situează în regiunea de non-respingere nu vom respinge H_0.
7) *Concluzia*: atunci când respingem ipoteza nulă H_0 o preferăm pe cea alternativă H_1 (ca urmare a depășirii nivelului de încredere $P = 1 - \alpha$ (de obicei 95% sau 99%). Când H_0 nu este respinsă, atunci concluzionăm că ipoteza nulă ar putea fi adevărată.

Așa cum am prezentat detaliat în capitolele anterioare, *eroarea standard* a unei statistici determinate la nivel de eșantion, care estimează de obicei un parametru al populației, reprezintă deviația standard a distribuției sale de sondaj. Este un element cheie în construirea intervalelor de încredere. Dacă statistica de interes este media eșantionului, atunci avem eroarea standard a mediei ($\sigma_{\bar{x}}$), care măsoară dispersia sau împrăștierea mediilor eșantioanelor de același volum n extrase dintr-o populație N în jurul mediei acestei populații: $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.

Scanned with OKEN Scanner

Valoarea p poate fi privită ca fiind *probabilitatea exactă* ca valoarea calculată a testului statistic la nivel de eșantion (z, t, χ^2, F, *etc.*) să fie cel puțin la fel de „neobișnuită" sau „extremă" ca valoarea teoretică, atunci când H_0 este adevărată (nu există niciun efect, nicio diferență sau nicio relație semnificativă între variabilele studiate). Cu alte cuvinte, dacă ipoteza nulă H_0 ar fi adevărată, cât de probabil este să vedem o valoare „atât de extremă" cum este valoarea testului nostru statistic, calculată pe baza datelor din eșantionul studiat? Valoarea p reprezintă cea mai mică valoare a nivelului de semnificație α, asociat testului statistic, pentru care putem respinge ipoteza nulă H_0.

Atunci când valoarea probabilității p este redusă, să presupunem sub nivelul de 0,05 sau 5%, rezultă faptul că rezultatul se obține destul de rar, în condițiile ipotezei noastre nule, ceea ce ridică semne de întrebare serioase și ne determină să luăm în considerare varianta că există șanse mari ca H_0 să nu fie adevărată. În consecință, dacă valoarea lui p este egală cu 0,05 înseamnă că este foarte puțin probabil să fi obținut acel rezultat doar din întâmplare, în condițiile în care H_0 este adevărată, și afirmăm că rezultatul nostru este *semnificativ statistic*. Avem astfel doar 5% șanse să fi obținut rezultatul nostru de la nivelul eșantionului în mod cu totul întâmplător, suntem aproape convinși că există un efect real, o diferență sau o relație între variabilele analizate și în concluzie respingem ipoteza nulă H_0.

La polul opus, dacă valoarea lui p este mare ($p > 0{,}05$), rezultatul obținut nu este semnificativ statistic și nu avem suficiente dovezi în vederea respingerii ipotezei nule H_0. Nerespingerea ipotezei nule nu semnifică faptul că ea este considerată adevărată, ci doar că nu avem suficiente dovezi furnizate de eșantionul nostru împotriva ei. Dacă obținem în practică $p = 0{,}35$ înseamnă că există 35% șanse ca rezultatul investigat (diferența, asocierea) să apară întâmplător, cu alte cuvinte „riscăm foarte mult" dacă respingem ipoteza nulă H_0.

În concluzie, nivelul probabilității p nu evidențiază faptul că H_0 este adevărată sau falsă, ci arată cât de compatibil este rezultatul cercetării cu ipoteza nulă. Să presupunem de exemplu, că dorim să testăm dacă un medicament nou scade tensiunea arterială mai mult comparativ cu placebo. Ipotezele studiului vor fi: H_0: medicamentul nu are efect (diferența medie este zero) și H_1: medicamentul are efect (diferența medie este diferită de zero). Presupunem că în urma aplicării testului statistic Student, obținem o diferență medie calculată la nivel de eșantion de $-6\ mmHg$ și o probabilitate asociată $p = 0{,}014$. Având în vedere faptul că $p < 0{,}05$, concluzionăm că rezultatul testului nostru este semnificativ statistic, existând o probabilitate foarte mică ca această diferență ($-6\ mmHg$) să fie obținută din întâmplare, atunci când medicamentul nu are niciun efect. Vom respinge astfel ipoteza nulă H_0 și vom afirma că eșantionul nostru furnizează suficiente dovezi că medicamentul are efect (reduce nivelul tensiunii arteriale în comparație cu placebo), în cazul populației generale din care acesta provine.

Scanned with OKEN Scanner

9.2. TESTAREA IPOTEZELOR PENTRU MEDIA UNEI POPULAȚII - TESTELE „z" ȘI „t"

În procesul de testare a ipotezelor statistice pentru media unei populații ne putem situa în practică într-un dintre următoarele două situații: prima, în situația când eșantionul provine dintr-o populație (aproximativ) normal distribuită sau dintr-o populație care nu urmează o distribuție normală dar volumul său $n \geq 30$ (teorema limită centrală), având varianța (σ^2) *cunoscută* și a doua când eșantionul provine dintr-o populație (aproximativ) normal distribuită având varianța (σ^2) *necunoscută*. Vom detalia și exemplifica în continuare ambele variante.

A. eșantionul provine dintr-o populație (aproximativ) normal distribuită sau dintr-o populație care nu urmează o distribuție normală dar volumul său $n \geq 30$, având varianța (σ^2) cunoscută

Testul statistic utilizat pentru a testa ipoteza nulă H_0: $\mu = \mu_0$ împotriva ipotezei alternative H_1: $\mu \neq \mu_0$ este:

$$z_{calculat} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \tag{9.2}$$

care, atunci când H_0 este adevărată, urmează o distribuție normală standard.

În cadrul distribuției de sondaj a mediilor eșantioanelor extrase dintr-o populație normal distribuită, $\bar{x}$ reprezintă un *estimator nedeplasat* al mediei populației μ, având varianța:

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} \tag{9.3}$$

Rezultă că, eroarea standard a mediei eșantionului va fi:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \tag{9.4}$$

Nivelul de semnificație $\alpha = 0{,}05$ va fi împărțit în mod egal între cele două cozi laterale $\alpha/2 = 0{,}025$ asociat cu valorile foarte mici și $\alpha/2 = 0{,}025$ asociat cu valorile foarte mari.

Valoarea critică $z_{1-\alpha/2} = 1{,}96$ reprezintă abscisa corespunzătoare probabilității cumulate $P = 1 - \frac{\alpha}{2} = 1 - \frac{0{,}05}{2} = 1 - 0{,}025 = 0{,}975$

Regiunea de respingere a testului z va fi în compusă din 2 părți:

$$z_{calculat} \leq -z_{1-\alpha/2} \text{ și } z_{calculat} \geq z_{1-\alpha/2}$$

Scanned with OKEN Scanner

Vom respinge H_0 dacă valoarea calculată a testului este cuprinsă în regiunea de respingere.

În practică, deosebim 3 situații distincte: test bilateral (regiunea de respingere este situată în ambele cozi ale distribuției) test unilateral dreapta / stânga (figura 9.1).

Exemplu: suntem interesați de cercetarea vârstei medii a unei anumite populații și formulăm următoarea întrebare: putem concluziona că vârsta medie a acestei populații este diferită de 45 de ani? Datele disponibile sunt reprezentate de un eșantion format din 10 indivizi, selectați aleatoriu din această populație, pentru care s-a calculat o vârstă medie de 42 de ani. Presupunem că eșantionul provine dintr-o populație care urmează o distribuție aproximativ normală, pentru care se cunoaște din cercetări anterioare valoarea variației $\sigma^2 = 20$.

Formulăm cele două ipoteze de lucru, în varianta testului *bilateral*:

- Ipoteza nulă H_0: vârsta medie a populației este 45 ani ($\mu = 45$)
- Ipoteza alternativă H_1: vârsta medie a populației nu este de 45 de ani ($\mu \neq 45$)

Considerăm nivelul de semnificație asociat testului: $\alpha = 0{,}05$, pentru care valoarea critică este: $z_{1-\alpha/2} = 1{,}96$.

Testul statistic utilizat pentru a testa ipoteza nulă H_0:

$$z_{calculat} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{42 - 45}{\sqrt{20/10}} = \frac{-3}{1{,}41} = -2{,}12$$

Testul statistic z este distribuit normal, având media 0 și varianța 1. În situația de față există un număr enorm de valori ale testului z, câte una pentru fiecare eșantion de volum $n = 10$ extras din această populație, din care noi dispunem doar de una, cea calculată pentru eșantionul nostru ($z_{calculat} = -2{,}12$). Construim în figura 9.2 distribuția teoretică a valorilor testelor z pentru toate eșantioanele de volum $n = 10$ extrase din populația generală, care este centrată în jurul valorii $z = 0$, în condițiile ipotezei nule H_0 adevărate (valoarea $z = 0$ se obține atunci când $\bar{x} = \mu_0$, adică vârsta medie a populației este de 45 ani). Cu cât valoarea $\bar{x}$ calculată pentru eșantionul nostru, în calitate de estimator punctual al mediei populației generale μ, este mai mare sau mai mică decât nivelul de 45 de ani, cu atât ne deplasăm în stânga către $-\infty$ sau în dreapta spre $+\infty$, pe axa absciselor din graficul distribuției normale a testului z (figura 9.2).

Întrebarea care se pune în acest caz este: cât de „extremă" sau „neobișnuită" ar trebui să fie valoarea testului statistic z, calculat la nivelul eșantionului studiat, pentru a se încadra în regiunea de respingere?

Vom respinge H_0, dacă valoarea calculată a testului este cuprinsă în regiunea de respingere, pentru $\alpha = 0{,}05$. Conform datelor furnizate de

Scanned with OKEN Scanner

eșantionul nostru avem $z_{calculat} = -2{,}12$, valoare situată în mod evident în regiunea de respingere (zona marcată cu roșu din figura 9.2).

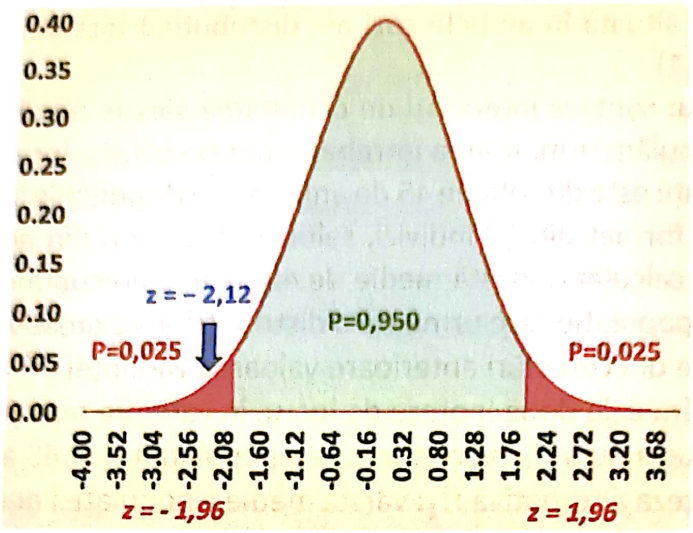


Figura 9.2. Distribuția valorilor testului statistic z bilateral pentru toate eșantioanele de volum $n = 10$

Prin urmare, vom respinge ipoteza nulă H_0 și vom afirma că valoarea calculată a testului este semnificativă statistic la un nivel $\alpha = 0{,}05$. Concluzionăm astfel că media populației studiate nu este egală cu 45 de ani și vom lua decizia finală conform acestui rezultat.

O alternativă des întâlnită în literatura de specialitate se referă la raportarea *probabilității exacte* p de a obține o valoare la fel de „neobișnuită" (extremă) comparativ cu cea calculată, în situația când ipoteza nulă H_0 este adevărată.

Tabelul 9.1. Distribuția normală standard în cazul testului bilateral

DISTRIBUTIA NORMALA STANDARD

z	Probabilitati cumulate						
	0.00	0.01	0.02	0.03	0.04	0.05	0.06
0.00	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239
0.10	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636
...	...	...	...	...	...	...	...
1.80	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686
1.90	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750
2.00	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803
2.10	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846

Scanned with OKEN Scanner

Pentru determinarea probabilității p se citește din tabelul distribuției normale standard (tabelul 9.1) probabilitatea cumulată aferentă valorii 2,12, care este $p = 0{,}983$. Astfel, când H_0 este adevărată, probabilitatea de a obține o valoare a lui z cel puțin la fel de mare ca 2,12 este $1 - p = 1 - 0{,}983 = 0{,}017$ și, fiind un test bilateral, probabilitatea de a observa o valoare a lui z cel puțin la fel de scăzută ca $-2{,}12$ este tot 0,017 (curba normală este simetrică). Raportarea valorii p asociate a testului statistic ne permite să evidențiem cât de rară este aceasta în situația ipotezei nule adevărate. În cazul exemplului nostru, avem $p = 0{,}034$ $(0{,}017 + 0{,}017)$, cu alte cuvinte probabilitatea de a obține o valoare cel puțin la fel de „extremă" cum este 2,12 în ambele direcții, când ipoteza nulă este adevărată este 0,034 (sau 3,4%). Se poate utiliza și în acest caz funcția Excel: $= NORM.S.DIST(-2{,}12; TRUE)$, care returnează valoarea 0,017.

Testele statistice pot fi însă și *unilaterale*, situație în care regiunea de respingere este situată doar într-una dintre cozile laterale ale distribuției testului statistic. Mergând pe linia exemplului anterior, în loc să investigăm dacă media poate fi considerată diferită de 45 de ani, cercetăm dacă putem concluziona că vârsta medie a populației este mai mică de 45 de ani (test unilateral stânga prezentat în figura 9.3). Astfel, avem:

- Ipoteza nulă H_0: vârsta medie a populației este mai mare sau egală cu 45 ani ($\mu \geq 45$)
- Ipoteza alternativă H_1: vârsta medie a populației este mai redusă de 45 de ani ($\mu < 45$)

Considerăm nivelul de semnificație asociat testului: $\alpha = 0{,}05$ Valoarea critică $z_{1-\alpha/2} = -1{,}645$ (valoarea critică a lui z, la stânga căreia se situează 0,05 sau 5% din aria de sub curba normală).

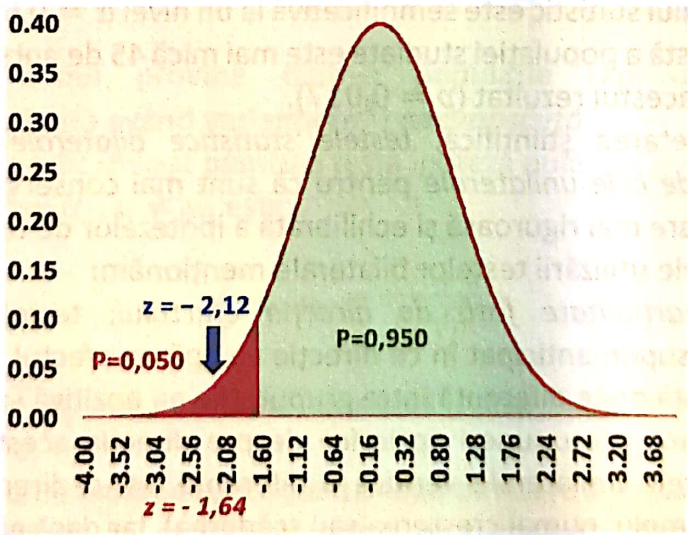


Figura 9.3. Distribuția valorilor testului statistic z unilateral stânga pentru toate eșantioanele de volum $n = 10$

Scanned with OKEN Scanner

Testul statistic utilizat pentru a verifica ipoteza nulă H_0:

$$z_{calculat} = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{42 - 45}{\sqrt{20/10}} = \frac{-3}{1{,}41} = -2{,}12$$

Vom respinge H_0 dacă valoarea calculată a testului este cuprinsă în regiunea de respingere $z_{calculat} \leq -1{,}645$, pentru $\alpha = 0{,}05$. Conform datelor din eșantionul nostru avem $z_{calculat} = -2{,}12 < -1{,}645$, valoare situată în regiunea de respingere.

Tabelul 9.2. Distribuția normală standard în cazul testului unilateral stânga

DISTRIBUTIA NORMALA STANDARD

z	Probabilitati cumulate					
	0.00	0.01	0.02	0.03	0.04	0.05
0.00	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199
0.10	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596
0.20	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987
...	...	...	...	...	...	...
1.50	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394
1.60	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505
1.70	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599

Prin urmare, vom respinge ipoteza nulă H_0 și vom afirma că valoarea calculată a testului statistic este semnificativă la un nivel $\alpha = 0{,}05$. Concluzionăm că media de vârstă a populației studiate este mai mică 45 de ani și vom lua decizia finală conform acestui rezultat ($p = 0{,}017$).

În cercetarea științifică, *testele statistice bilaterale sunt de obicei preferate față de cele unilaterale* pentru că sunt mai conservatoare și neutre, oferind o evaluare mai riguroasă și echilibrată a ipotezelor de testat.

Printre avantajele utilizării testelor bilaterale menționăm:

- *imparțialitate față de direcția efectului*: testele bilaterale nu presupun anticipat în ce direcție va apărea efectul. Ele verifică dacă există orice diferență între grupuri (fie ea pozitivă sau negativă), fără a face presupuneri apriorice despre direcția acesteia. În contrast, testele unilaterale verifică o diferență într-o direcție specifică (de exemplu, numai creșterea sau scăderea), iar dacă efectul real este în direcția opusă, acest lucru ar putea să nu fie detectat;
- *risc mai mic de a face erori de tip I*: în testele unilaterale, ipoteza nulă este respinsă doar dacă rezultatul este suficient de extrem într-o

Scanned with OKEN Scanner

anumită direcție (doar o creștere sau doar o scădere). Astfel, probabilitatea erorii de tip I $\alpha = 0{,}05$ este concentrată într-o singură coadă a distribuției (pozitivă sau coada negativă). Spre deosebire de testele unilaterale, în testele bilaterale, ipoteza nulă presupune că nu există nicio diferență între grupuri sau variabile, iar noi suntem interesați să vedem dacă datele arată o diferență semnificativă în orice direcție. Acest lucru înseamnă că trebuie să verificăm ambele "cozi" ale distribuției, atât valorile extreme pozitive, cât și cele extreme negative (la un test bilateral, acest prag $\alpha = 0{,}05$ va fi împărțit în mod egal între ambele părți ale distribuției: $0{,}025$ pentru coada din stânga și $0{,}025$ pentru coada din dreapta);

- *neutralitate și validitate în raport cu ipoteza nulă*: într-o abordare științifică riguroasă, este important să se adopte o poziție neutră, investigând dacă există o diferență semnificativă, fără a presupune de la început că aceasta va avea o anumită direcție;
- *acceptabilitate mai mare în mediul academic*: comunitatea științifică este adesea sceptică față de testele unilaterale deoarece acestea pot fi văzute ca o formă de manipulare a rezultatelor pentru a obține mai ușor o semnificație statistică: "$p - hacking$" (un test bilateral are standarde mai stricte și este mai acceptat ca metodă robustă și corectă);
- *detectarea efectelor neașteptate*: în anumite situații, efectele într-o direcție opusă față de cea anticipată ar putea fi foarte importante din punct de vedere științific. Un test bilateral este capabil să detecteze aceste efecte neașteptate, ceea ce nu se întâmplă în cazul unui test unilateral, care este limitat la o singură direcție a ipotezei.

B. eșantionul provine dintr-o populație (aproximativ) normal distribuită având varianța (σ^2) *necunoscută*

Testul statistic utilizat pentru a testa ipoteza nulă H_0: $\mu = \mu_0$ împotriva ipotezei alternative H_1: $\mu \neq \mu_0$ este:

$$t_{calculat} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{9.5}$$

Care, atunci când H_0 este adevărată, urmează o distribuție Student (t) cu $n - 1$ grade de libertate.

În situația când nu se cunoaște deviația standard a populației generale se utilizează, în calitate de estimator al acesteia, deviația standard ajustată s calculată la nivel de eșantion:

Scanned with OKEN Scanner

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \quad (9.6)$$

Nivelul de semnificație $\alpha = 0{,}05$ va fi împărțit în mod egal între cele două cozi laterale $\alpha/2 = 0{,}025$ asociat cu valorile foarte mici și $\alpha/2 = 0{,}025$ asociat cu valorile foarte mari. Pentru determinarea valorii critice $t_{n-1,\ 1-\alpha/2}$ putem folosi tabelul distribuției Student sau funcția Excel $= T.INV(0{,}975; n-1)$.

Regiunea de respingere a testului t va fi în consecință compusă din 2 părți:

$$t_{calculat} \leq -t_{n-1,\ 1-\alpha/2} \quad \text{și} \quad t_{calculat} \geq t_{n-1,\ 1-\alpha/2}$$

Exemplu: Au fost înregistrate date privind presiunea intraoculară (mm Hg) pentru un eșantion format din 21 de subiecți, cu vârsta peste 55 de ani. Putem concluziona în baza acestor date că media presiunii intraoculare la nivelul populației din care a fost selectat acest eșantion este diferită de 14? Stabilim nivelul de semnificație $\alpha = 0{,}05$.

Tabelul 9.3 Presiunea intraoculară (mm Hg) a celor 21 de subiecți din eșantionul de lucru

14,5	12,9	17,0
12,9	17,9	10,0
14,0	12,0	18,5
16,1	16,4	20,8
12,0	24,2	16,2
17,5	12,2	14,9
14,1	14,4	19,6

Considerăm că cei 21 de subiecți reprezintă un eșantion extras în mod aleatoriu dintr-o populație de persoane având vârstele peste 55 de ani. De asemenea, presupunem că presiunea intraoculară în cadrul populației generale, din care provine eșantionul, urmează o distribuție aproximativ normală.

Formulăm ipotezele statistice ale studiului:

- Ipoteza nulă H_0: presiunea intraoculară medie a populației este 14 mm Hg ($\mu = 14$)
- Ipoteza alternativă H_1: presiunea intraoculară medie a populației nu este de 14 ($\mu \neq 14$)

Nivelul de semnificație asociat testului stabilit anterior demarării cercetării este $\alpha = 0{,}05$. Deoarece varianța populației generale este necunoscută, testul statistic utilizat urmează pentru a verifica ipoteza nulă

Scanned with OKEN Scanner

(H_0: $\mu = 14$) o distribuție Student cu $n - 1 = 21 - 1 = 20$ grade de libertate. Avem un test bilateral și prin urmare, $\alpha = 0{,}05$ va fi împărțit în mod egal între cele două cozi laterale $\alpha/2 = 0{,}025$ asociat cu valorile foarte mici și $\alpha/2 = 0{,}025$ asociat cu valorile foarte mari.

Așa cum se observă în tabelul 9.4, valoarea critică a testului, citită din tabelul distribuției Student "t" (sau utilizând formula $= T.INV(0{,}975; 20)$ din Excel) este: $t_{n-1,\ 1-\alpha/2} = 2{,}086$

Tabelul 9.4. Distribuția Student în cazul unui test bilateral

DISTRIBUȚIA STUDENT "t"

Grade de libertate	Suprafața delimitată în extremitatea dreaptă				
	0.10	0.05	0.025	0.01	0.005
1.00	3.0777	6.3138	12.7062	31.8205	63.6567
2.00	1.8856	2.9200	4.3027	6.9646	9.9248
3.00	1.6377	2.3534	3.1824	4.5407	5.8409
4.00	1.5332	2.1318	2.7764	3.7469	4.6041
...	...	...	...	...	...
19.00	1.3277	1.7291	2.0930	2.5395	2.8609
20.00	1.3253	1.7247	2.0860	2.5280	2.8453

Presiunea intraoculară medie, calculată la nivel de eșantion:

$$\bar{x} = \frac{\sum_{i=1}^{21} x_i}{n} = \frac{328{,}10}{21} = 15{,}62 \text{ mm Hg}$$

Deviația standard ajustată calculată la nivelul eșantionului pentru a estima parametrul corespunzător σ necunoscut al populației:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{15{,}13}{21-1}} = 3{,}38 \text{ mm Hg}$$

În final, determinăm valoarea testului statistic utilizat pentru a testa ipoteza nulă H_0:

$$t_{calculat} = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{15{,}62 - 14{,}00}{3.38/\sqrt{21}} = 2{,}20$$

Scanned with OKEN Scanner

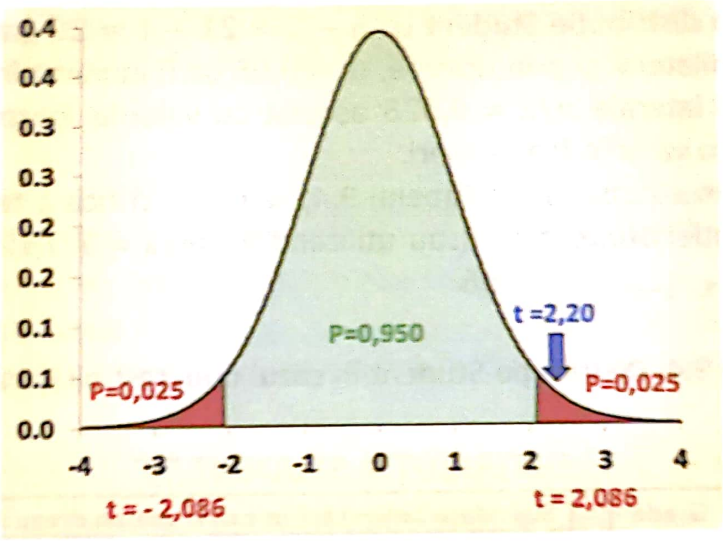


Figura 9.4. Distribuția valorilor testului statistic t bilateral pentru toate eșantioanele de volum $n = 21$

Vom respinge ipoteza nulă H_0 dacă valoarea calculată a testului este cuprinsă în regiunea de respingere: $t_{calculat} \leq -2{,}086$ și $t_{calculat} \geq 2{,}086$, pentru $\alpha = 0{,}05$. Conform datelor din eșantionul nostru avem $t_{calculat} = 2{,}20$, valoare situată în regiunea de respingere (figura 9.4).

Prin urmare, vom respinge ipoteza nulă H_0 și vom afirma că valoarea calculată a testului este semnificativă statistic la un nivel $\alpha = 0{,}05$. Concluzionăm că media populației studiate nu este egală cu 14 mm Hg.

Prezentăm în continuare etapele derulării testului „t", pentru același exemplu, în programul SPSS:

- Calculăm *indicatorii statistici descriptivi*, determinați la nivel de eșantion:

Descriptives

			Statistic	Std. Error
Presiune_intraoculara	Mean		15.6238	.73821
	95% Confidence Interval for Mean	Lower Bound	14.0839	
		Upper Bound	17.1637	
	5% Trimmed Mean		15.4635	
	Median		14.9000	
	Variance		11.444	
	Std. Deviation		3.38288	
	Minimum		10.00	
	Maximum		24.20	
	Range		14.20	
	Interquartile Range		4.80	
	Skewness		.732	.501
	Kurtosis		.668	.972

Scanned with OKEN Scanner

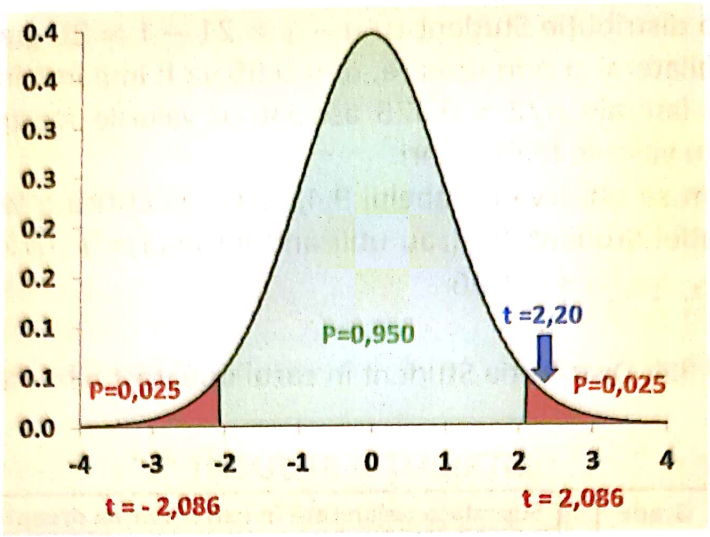


Figura 9.4. Distribuția valorilor testului statistic t bilateral pentru toate eșantioanele de volum $n = 21$

Vom respinge ipoteza nulă H_0 dacă valoarea calculată a testului este cuprinsă în regiunea de respingere: $t_{calculat} \leq -2{,}086$ și $t_{calculat} \geq 2{,}086$, pentru $\alpha = 0{,}05$. Conform datelor din eșantionul nostru avem $t_{calculat} = 2{,}20$, valoare situată în regiunea de respingere (figura 9.4).

Prin urmare, vom respinge ipoteza nulă H_0 și vom afirma că valoarea calculată a testului este semnificativă statistic la un nivel $\alpha = 0{,}05$. Concluzionăm că media populației studiate nu este egală cu 14 mm Hg.

Prezentăm în continuare etapele derulării testului „t", pentru același exemplu, în programul SPSS:

- Calculăm *indicatorii statistici descriptivi*, determinați la nivel de eșantion:

Descriptives

			Statistic	Std. Error
Presiune_intraoculara	Mean		15.6238	.73821
	95% Confidence Interval for Mean	Lower Bound	14.0839	
		Upper Bound	17.1637	
	5% Trimmed Mean		15.4635	
	Median		14.9000	
	Variance		11.444	
	Std. Deviation		3.38288	
	Minimum		10.00	
	Maximum		24.20	
	Range		14.20	
	Interquartile Range		4.80	
	Skewness		.732	.501
	Kurtosis		.668	.972

Scanned with OKEN Scanner

Realizăm o primă evaluare a normalității distribuției valorilor presiunii intraoculare, inițial prin intermediul indicatorilor de statistică descriptivă, precum media, mediana, deviația standard, coeficienții de asimetrie și de boltire.

Media și mediana au valori apropiate, sau chiar identice în cazul distribuției normale. În situația datelor din eșantionul nostru, observăm o apropiere a valorilor celor doi indicatori ai tendinței centrale ($\bar{x} = 15{,}6$ și $Me = 14{,}90$). De asemenea se poate compara media cu deviația standard, deoarece în eșantioanele care provin din populații normal distribuite media este superioară deviației standard. Și acest criteriu este îndeplinit ($\bar{x} = 15{,}62$ și $\sigma = 3{,}38$), media fiind de peste 4,6 ori mai mare comparativ cu deviația standard din eșantion.

Coeficienții de asimetrie și boltire trebuie să înregistreze valori apropiate de zero, valorile care nu se încadrează în intervalul $(-1{,}1)$ sugerând abateri de la normalitate (asimetrie la stânga/dreapta și/sau aplatizare/boltire mai mari decât în cazul distribuției gausiene). În cazul de față, coeficientul de asimetrie este 0,732 și coeficientul de boltire este 0,668, ambele valori sugerând abateri mici de la normalitate.

- Realizăm în continuare *histograma* seriei, pentru a observa vizual distribuția de frecvență a variabilei „presiunea intraoculară".

În situația în care histograma are o formă de clopot și este aproximativ simetrică ($\bar{x} = Me$), aceasta sugerează normalitatea distribuției variabilei statistice analizate (figura 9.5):

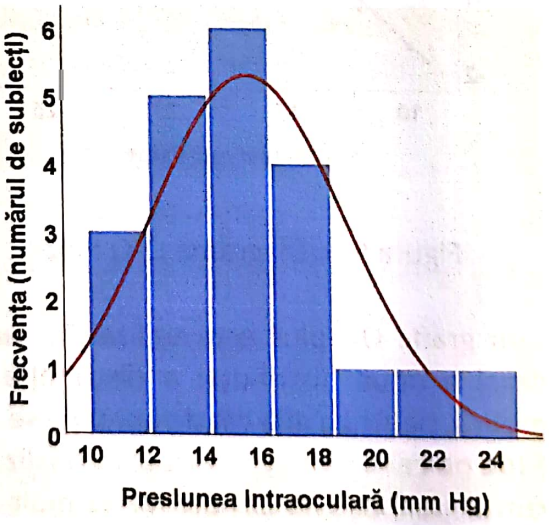


Figura 9.5. Distribuția variantelor variabilei „presiunea intraoculară" la nivelul eșantionului de volum $n = 21$

Observăm din figura 9.5 o tendință de concentrare a variantelor variabilei „presiunea intraoculară" în jurul tendinței centrale, reprezentată de media

Scanned with OKEN Scanner

aritmetică a seriei, și o reducere a numărului de cazuri pe măsură ce ne apropiem de valorile extreme, în ambele direcții.

- Construim *diagrama Q-Q Plot*, permite cel mai bine evaluarea normalității distribuției unei populații, întrucât observăm toate valorile din seria de date, reprezentate prin puncte.

Graficul poate fi utilizat pentru a stabili din punct de vedere vizual dacă un set de date provine dintr-o anumită distribuție statistică (de obicei cea normală). În acest scop datele din eșantion sunt sortate crescător și apoi reprezentate grafic împreună cu cuantilele corespunzătoare, calculate pentru o distribuție normală. În situația în care punctele se vor situa foarte aproape de linia de referință înclinată la 45^o, distribuția datelor din populația studiată este aproximativ normală. Atunci când punctele deviază substanțial de la linia de referință, datele nu sunt normal distribuite (figura 9.6).

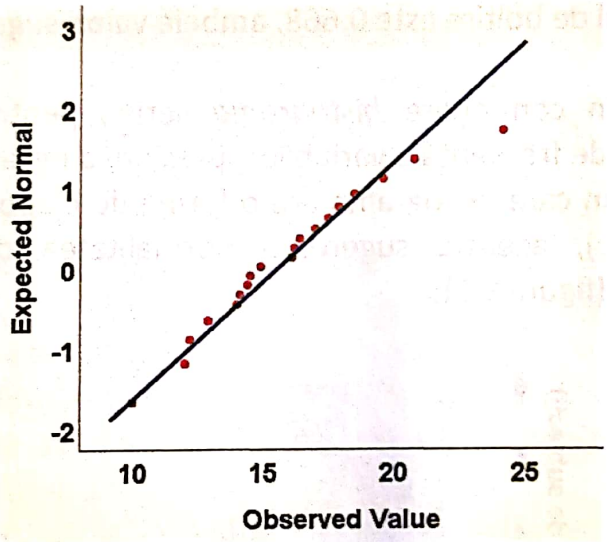


Figura 9.6. Diagrama Q-Q Plot

Sintetizând, un grafic QQ-plot este utilizat în vederea comparării din punct de vedere vizual a două distribuții: o distribuție a datelor observate (empirică) și una teoretică, pentru a afla cât de aproape se situează cea teoretică față de cea empirică (de obicei distribuția normală). Analizând datele prezentate în figura 9.6, observăm că punctele nu deviază foarte mult de la linia de referință și în consecință putem concluziona că eșantionul nostru provine dintr-o populație care urmează o distribuție aproximativ normală.

- O altă metodă de a evalua normalitatea distribuțiilor populațiilor din care provin eșantioanele de studiu este utilizarea *testelor statistice de normalitate*.

Scanned with OKEN Scanner

Există numeroase teste specifice pentru normalitate: Shapiro-Wilk (folosit mai ales pentru eșantioane având sub 50 de subiecți), Kolmogorov Smirnov, D'Agostino-Pearson și testul Chi-pătrat.

Testele specifice de normalitate (Shapiro-Wilk, Kolmogorov Smirnov) compară scorurile calculate la nivel de eșantion cu un set de scoruri normal distribuit, având aceeași medie și deviație standard. Ipoteza nulă H_0 reprezintă situația în care nu există diferențe semnificative statistic între distribuția empirică a datelor și distribuția normală, respectiv ipoteza alternativă H_1 reprezintă situația în care există diferențe semnificative statistic între distribuția datelor și distribuția normală.

În funcție de valoarea probabilității p aferentă testului de normalitate utilizat vom decide dacă:

- ✓ respingem ipoteza nulă H_0 a unei populații normal distribuite ($p < 0{,}05$);
- ✓ nu respingem ipoteza nulă H_0 și considerăm că populația din care a fost extras în mod aleatoriu eșantionul studiat ar putea fi normal distribuită ($p \geq 0{,}05$).

Problema acestor teste este că, în situația eșantioanelor de volum redus au putere scăzută de a respinge ipoteza nulă a normalității distribuției și, în acest caz, eșantioanele mici de multe ori trec testele de normalitate. La polul opus, în cazul eșantioanelor de volum mare, testele „sunt prea puternice" și se vor obține rezultate semnificative statistic (va fi respinsă ipoteza nulă a normalității) și atunci când deviațiile de la normalitate ale distribuției sunt destul de mici.

Generăm în programul SPSS testele de verificare a normalității datelor din populația generală, din care a fost extras eșantionul de 21 de subiecți:

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Presiune_intraoculara	.109	21	.200*	.964	21	.594

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Întrucât în situația ambelor teste de verificare a normalității valoarea probabilității $p \geq 0{,}05$, concluzionăm că populația din care provine eșantionul studiat ar putea fi normal distribuită.

- ➢ Ultima etapă constă în calcularea testului Student „t" pentru compararea mediei presiunii intraoculare din populația generală cu nivelul de referință 14 ($\mu = 14$):

Scanned with OKEN Scanner

One-Sample Test

Test Value = 14

	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Presiune_intraoculara	2.200	20	.040	1.62381	.0839	3.1637

Având în vedere probabilitatea aferentă testului Student aplicat $p = 0{,}04$ este mai redusă decât nivelul de semnificație $\alpha = 0{,}05$, vom respinge ipoteza nulă H_0 și vom afirma că valoarea calculată a testului este semnificativă statistic la un nivel $\alpha = 0{,}05$. Concluzionăm că media populației studiate nu este egală cu 14 mm Hg. Observăm faptul că, programul SPSS calculează și intervalul de încredere 95% pentru diferența de 1,62 dintre valoarea calculată la nivel de eșantion în calitate de estimator punctual al presiunii intraoculare medii a populației (15,62 mm Hg) și valoarea de referință (14,00 mm Hg). Intervalul de încredere astfel determinat: (0,0839; 3,1637) nu include valoarea zero și prin urmare considerăm că media presiunii intraoculare de la nivel populațional este semnificativ diferită de nivelul de referință (14,00 mm Hg).

Indiferent că am aplicat testul Student de verificare a ipotezei $H_0: \mu = 14$ sau am calculat intervalul de încredere pentru diferența dintre media populației și valoarea de referință, rezultatele converg către aceeași concluzie: media presiunii intraoculare a populației generale, din care a fost extras în mod aleatoriu eșantionul studiat, este semnificativ diferită de 14 mm Hg.

9.3. TESTAREA IPOTEZELOR PENTRU PROPORȚIA UNEI POPULAȚII - TESTUL „z"

Este similară cu procedura anterioară realizată în cazul mediei. Atunci când volumul eșantionului este suficient de mare ($n \cdot p \geq 5$ și $n \cdot (1-p) \geq 5$) se aplică Teorema Limită Centrală.

Testul statistic utilizat pentru a testa ipoteza nulă: $H_0: p = p_0$ ($H_1: p \neq p_0$) este:

$$z_{calculat} = \frac{p - p_0}{\sqrt{\frac{p_0 \cdot (1-p_0)}{n}}} \tag{9.7}$$

care, atunci când H_0 este adevărată, urmează o distribuție normală standard.

Nivelul de semnificație $\alpha = 0{,}05$ va fi împărțit în mod egal între cele două cozi laterale $\alpha/2 = 0{,}025$ asociat cu valorile foarte mici și $\alpha/2 = 0{,}025$ asociat cu valorile foarte mari.

Scanned with OKEN Scanner

Valoarea critică $z_{1-\alpha/2} = 1{,}96$ reprezintă abscisa corespunzătoare probabilității cumulate $P = 1 - \alpha/2 = 1 - 0{,}05/2 = 1 - 0{,}025 = 0{,}975$. Ca alternativă, putem utiliza funcția Excel: $= NORM.S.INV(0{,}975)$.

Regiunea de respingere a testului z va fi în consecință compusă din 2 părți:

$$z_{calculat} \leq -z_{1-\alpha/2} \quad \text{și} \quad z_{calculat} \geq z_{1-\alpha/2}$$

Exemplu: presupunem că dorim să verificăm dacă rata complicațiilor post-operatorii s-a redus semnificativ după implementarea unui nou protocol de igienă într-un anumit spital. Înainte de implementarea noului protocol, rata complicațiilor infecțioase post-operatorii era de 12%. După ce protocolul a fost aplicat timp de trei luni, un spital vrea să verifice dacă proporția pacienților cu complicații a scăzut. Vom selecta în mod aleatoriu un eșantion de 250 de pacienți care au suferit intervenții chirurgicale ulterior implementării noului protocol. Dintre aceștia se constată faptul că un număr de 18 pacienți au înregistrat complicații post-operatorii. Echipa de management dorește să cunoască dacă proporția pacienților cu complicații post-operatorii este diferită în comparație cu nivelul anterior de 12%. Putem concluziona că proporția pacienților cu complicații post-operatorii după implementarea protocolului de igienă este diferită de 0,12? Cu alte cuvinte, este proporția observată la nivelul eșantionului extras în mod aleatoriu $p = 18/250 = 0{,}072$ „suficient" de diferită de valoarea de 0,12?

Putem aplica Teorema Limită Centrală deoarece $n \cdot p = 18 > 5$ și $n \cdot (1-p) = 232 > 5$ și în consecință distribuția binomială poate fi aproximată printr-o distribuție normală.

Formulăm ipotezele statistice ale studiului:

- Ipoteza nulă H_0: proporția pacienților cu complicații post-operatorii după implementarea protocolului de igienă este 0,12 ($p_0 = 0{,}12$)
- Ipoteza alternativă H_1: proporția pacienților cu complicații post-operatorii nu este 0,12 ($p_0 \neq 0{,}12$)

Considerăm nivelul de semnificație asociat testului: $\alpha = 0{,}05$. Avem un test bilateral: $\alpha = 0{,}05$ va fi împărțit în mod egal între cele două cozi laterale $\alpha/2 = 0{,}025$ asociat cu valorile foarte mici și $\alpha/2 = 0{,}025$ asociat cu valorile foarte mari.

Valoarea critică $z_{1-\alpha/2} = 1{,}96$ reprezintă abscisa corespunzătoare probabilității cumulate $P = 1 - \alpha/2 = 1 - 0{,}05/2 = 1 - 0{,}025 = 0{,}975$

Testul statistic utilizat pentru a testa ipoteza nulă $H_0: p = p_0$ este:

$$z_{calculat} = \frac{p - p_0}{\sqrt{\frac{p_0 \cdot (1 - p_0)}{n}}} = \frac{0{,}072 - 0{,}12}{\sqrt{\frac{0{,}12 \cdot (1 - 0{,}12)}{250}}} = -2{,}34$$

Scanned with OKEN Scanner

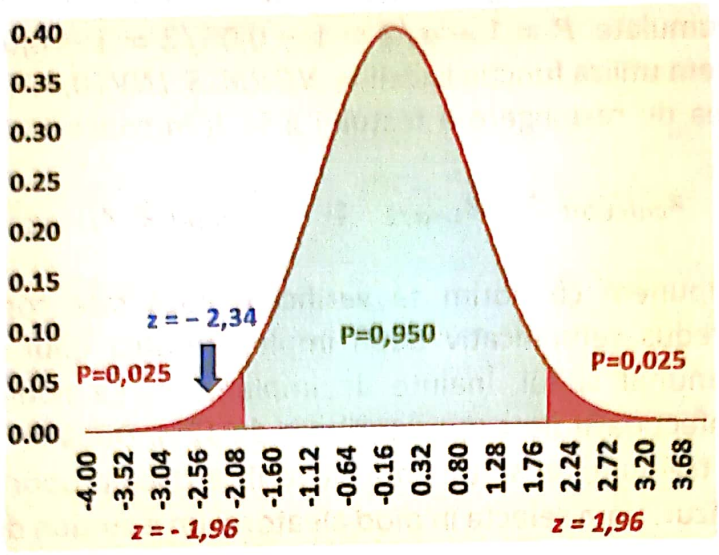


Figura 9.7. Distribuția valorilor testului statistic z bilateral pentru toate eșantioanele de volum $n = 250$

Presupunând că proporția pacienților cu complicații post-operatorii după implementarea protocolului de igienă în populația generală este 0,12 (H_0 este adevărată), proporția calculată la nivelul eșantionului de 0,072 este cu 2,34 deviații standard mai mică, fiind evident în zona de respingere.

Prin urmare, vom respinge ipoteza nulă H_0 și vom afirma că valoarea calculată a testului statistic calculat z este semnificativă la un nivel $\alpha = 0{,}05$. Concluzionăm că proporția pacienților cu complicații post-operatorii din populația studiată (toți pacienții care vor fi internați ulterior implementării protocolului de igienă) nu este egală cu 0,12 și vom lua decizia finală conform acestui rezultat.

Se citește din tabelul distribuției normale standard probabilitatea cumulată aferentă valorii calculate a testului $z_{calc} = -2{,}34$, care este $p = 0{,}0096$. În Excel putem folosi funcția:= $NORM.S.DIST(-2{,}34; TRUE)$. Astfel, când H_0 este adevărată, probabilitatea de a obține o valoare a lui z cel puțin la fel de redusă ca $-2{,}34$ este 0,0096 și, fiind un test bilateral, probabilitatea de a observa o valoare a lui z cel puțin la fel de ridicată ca 2,34 este tot 0,0096 (curba normală este simetrică).

Astfel, în cazul exemplului nostru, avem $p = 0{,}0096 + 0{,}0096 = 0{,}0192$. Cu alte cuvinte probabilitatea de a obține o valoare cel puțin la fel de „extremă" cum este 2,34 în ambele direcții, când ipoteza nulă este adevărată este 0,0192 (sau 1,92%).

Construim și intervalul de încredere 95% pentru proporția populației generale, după implementarea protocolului de igienă, utilizând formula 7.31, introdusă anterior în capitolul 7:

Scanned with OKEN Scanner

$$IC95\% = p \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{p \cdot (1-p)}{n}} = 0{,}072 \pm 1{,}96 \cdot \sqrt{\frac{0{,}072 \cdot 0{,}928}{250}}$$

Efectuând calculele, obținem: $IC95\% \approx (0{,}040; 0{,}104)$

În concluzie, noul protocol de igienă implementat în spital pare a fi eficient, conducând la o reducere a complicațiilor post-operatorii, cu o proporție estimată între 4,0% și 10,4% (nivel de încredere de 95%) în comparație situația existentă anterior, când procentul era mai ridicat, respectiv 12%.

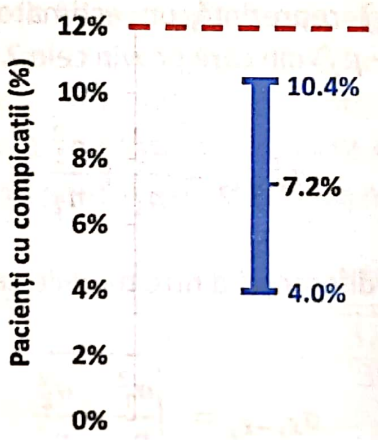


Figura 9.8. Intervalul de încredere 95% pentru proporția populației generale, după implementarea protocolului de igienă

Intervalul de încredere 95% construit în figura 9.8 ne arată proporția de pacienți cu complicații post-operatorii observată după implementarea noului protocol de igienă, evidențiind în mod clar incertitudinea estimării noastre, în jurul valorii determinate la nivel de eșantion de 7,2%. Putem astfel afirma faptul că, proporția reală a complicațiilor se situează probabil undeva între 4,0% și 10,4%, evident sub nivelul de 12%, existent anterior implementării protocolului.

9.4. TESTAREA IPOTEZELOR PENTRU DIFERENȚA MEDIILOR A DOUĂ POPULAȚII - TESTELE „z" ȘI „t"

În cadrul procesului de testare a ipotezelor statistice pentru diferența mediilor a două populații deosebim patru cazuri principale:

A. eșantioanele sunt independente și provin din 2 populații normal distribuite sau din populații care nu urmează o distribuție normală dar volumul lor $n \geq 30$ (teorema limită centrală) având varianțele (σ_1^2, σ_2^2) cunoscute

Scanned with OKEN Scanner

Testul statistic utilizat pentru a testa ipoteza nulă H_0: $\mu_1 = \mu_2$ sau $\mu_1 - \mu_2 = 0$ (H_1: $\mu_1 \neq \mu_2$ sau $\mu_1 - \mu_2 \neq 0$) este:

$$z_{calculat} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \quad (9.8)$$

care, atunci când H_0 este adevărată, urmează o distribuție normală standard.

În cadrul distribuției de sondaj a diferenței mediilor eșantioanelor extrase din 2 populații, $(\bar{x}_1 - \bar{x}_2)$ reprezintă un estimator nedeplasat al diferenței mediilor populațiilor $(\mu_1 - \mu_2)$ din care provin cele 2 eșantioane, având varianța:

$$\sigma^2_{\bar{x}_1 - \bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \quad (9.9)$$

Astfel, eroarea standard a diferenței dintre mediile celor 2 eșantioane va fi:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad (9.10)$$

Nivelul de semnificație $\alpha = 0{,}05$ va fi împărțit în mod egal între cele două cozi laterale $\alpha/2 = 0{,}025$ asociat cu valorile foarte mici și $\alpha/2 = 0{,}025$ asociat cu valorile foarte mari.

Valoarea critică $z_{1-\alpha/2} = 1{,}96$ reprezintă abscisa corespunzătoare probabilității cumulate $P = 1 - \alpha/2 = 1 - 0{,}05/2 = 1 - 0{,}025 = 0{,}975$ Se poate folosi și funcția Excel: $= NORM.S.INV(0{,}975)$.

Regiunea de respingere a testului z va fi în compusă din 2 părți:

$$z_{calculat} \leq -z_{1-\alpha/2} \quad \text{și} \quad z_{calculat} \geq z_{1-\alpha/2}$$

Exemplu: pentru a evalua efectul unui supliment alimentar asupra valorii glicemiei a jeun, la pacienții cu diabet zaharat de tip 2, au fost comparate mediile a două grupuri (eșantioane) independente (fiecare pacient aparține unui singur grup): un grup care a primit suplimentul ($n_1 = 30$) și un grup control care a primit placebo ($n_2 = 30$).

Valorile medii ale glicemiei au fost:

- grupul placebo: $\bar{x}_1 = 145$ mg/dL
- grupul supliment: $\bar{x}_2 = 130$ mg/dL

Scanned with OKEN Scanner

Presupunând că varianțele populațiilor, aproximativ normal distribuite, din care au fost extrase în mod aleatoriu aceste două eșantioane sunt cunoscute anterior demarării studiului ($\sigma_1^2 = 400$ și $\sigma_2^2 = 324$), vom utiliza în cadrul analizei testul z pentru diferența dintre două medii.

Formulăm ipotezele statistice ale studiului:

- Ipoteza nulă H_0: la finalul tratamentului, media glicemiei celor 2 populații este aceeași - suplimentul nu are efect ($\mu_1 = \mu_2$)
- Ipoteza alternativă H_1: la finalul tratamentului, media glicemiei celor 2 populații nu este aceeași - suplimentul are efect ($\mu_1 \neq \mu_2$)

Considerăm nivelul de semnificație asociat testului: $\alpha = 0{,}05$. Avem un test bilateral: α va fi împărțit în mod egal între cele două cozi laterale $\alpha/2 = 0{,}025$ asociat cu valorile foarte mici și $\alpha/2 = 0{,}025$ cu valorile foarte mari.

Valoarea critică $z_{1-\alpha/2} = 1{,}96$ reprezintă abscisa corespunzătoare probabilității cumulate $P = 1 - \alpha/2 = 1 - 0{,}05/2 = 1 - 0{,}025 = 0{,}975$

Testul statistic utilizat pentru a testa ipoteza nulă H_0: $\mu_1 = \mu_2$ este:

$$z_{calculat} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{145 - 130}{\sqrt{\frac{400}{30} + \frac{324}{30}}} = 3{,}05$$

Vom respinge H_0, dacă valoarea calculată a testului este cuprinsă în regiunea de respingere, pentru nivelul de semnificație dorit $\alpha = 0{,}05$. Conform datelor furnizate de eșantionul nostru avem $z_{calculat} = 3{,}05$, valoare situată în mod evident în regiunea de respingere (zona marcată cu roșu din figura 9.9).

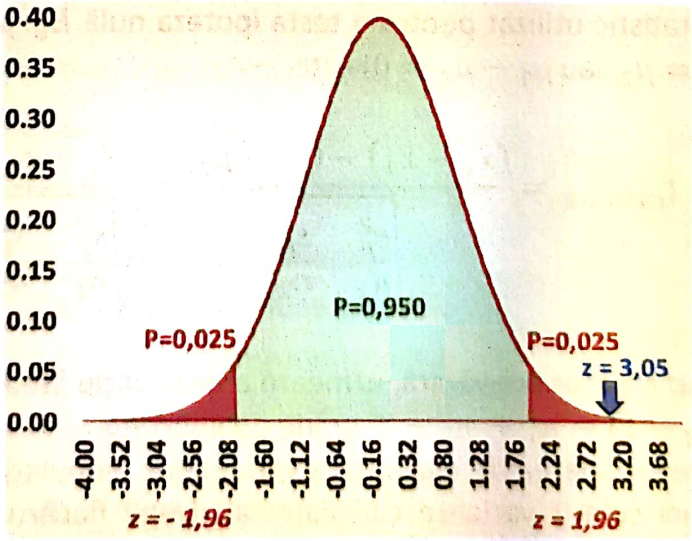


Figura 9.9. Distribuția valorilor testului statistic z bilateral pentru toate perechile de eșantioane extrase

Scanned with OKEN Scanner

În consecință, vom respinge ipoteza nulă H_0 și vom afirma că valoarea calculată a testului statistic calculat z este semnificativă la un nivel $\alpha = 0{,}05$. Rezultatele testului sugerează că există o diferență semnificativă statistic între cele două populații, în ceea ce privește nivelul glicemiei.

Pentru determinarea probabilității p se citește din tabelul distribuției normale standard probabilitatea cumulată aferentă valorii $z_{calculat} = 3{,}05$ care este $p = 0.999$. Astfel, când H_0 este adevărată, probabilitatea de a obține o valoare a lui z cel puțin la fel de mare ca $3{,}05$ este $1 - p = 1 - 0{,}999 = 0{,}001$ și, fiind un test bilateral, probabilitatea de a observa o valoare a lui z cel puțin la fel de scăzută ca $-3{,}05$ este tot $0{,}001$ (curba normală este simetrică). Raportarea valorii p asociate a testului statistic ne permite să evidențiem cât de rară este aceasta în situația ipotezei nule adevărate. În cazul exemplului nostru, avem $p = 0{,}002$ $(0{,}001 + 0{,}001)$, cu alte cuvinte probabilitatea de a obține o valoare cel puțin la fel de „extremă" cum este $3{,}05$ în ambele direcții, când ipoteza nulă este adevărată este $0{,}002$ (sau $0{,}2\%$). Se poate utiliza și în acest caz funcția Excel: $= NORM.S.DIST(3{,}05; TRUE)$, care returnează valoarea $0{,}999$.

Concluzionând, putem afirma faptul că, eșantioanele studiate furnizează dovezi în sprijinul ideii că administrarea suplimentului alimentar a determinat o scădere semnificativă a glicemiei față de placebo ($z = 3{,}05$, $p = 0{,}002 < 0{,}01$). Acest rezultat indică faptul că suplimentul administrat ar putea avea un efect benefic asupra controlului glicemic în cazul pacienților cu diabet zaharat de tip 2, la nivel populațional.

B. eșantioanele sunt independente și provin din 2 populații aproximativ normal distribuite, având variațele (σ_1^2, σ_2^2) necunoscute și egale

Testul statistic utilizat pentru a testa ipoteza nulă H_0: $\mu_1 = \mu_2$ sau $\mu_1 - \mu_2 = 0$ (H_1: $\mu_1 \neq \mu_2$ sau $\mu_1 - \mu_2 \neq 0$) este:

$$t_{calculat} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \tag{9.11}$$

care, atunci când H_0 este adevărată, urmează o distribuție Student (t) cu $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$ grade de libertate.

Dacă presupunerea de egalitate a variațelor populațiilor studiate este justificată, atunci cele 2 variațe calculate la nivelul fiecărui eșantion pot fi considerate drept estimatori ai unei variațe comune (s_p^2), determinată ca o medie ponderată a variațelor celor 2 eșantioane (s_1^2 și s_2^2). Ponderea fiecărei

Scanned with OKEN Scanner

varianțe de eșantion este reprezentată de gradele sale de libertate $(n_1 - 1)$ și $(n_2 - 1)$.

$$s_p^2 = \frac{s_1^2 \cdot (n_1 - 1) + s_2^2 \cdot (n_2 - 1)}{(n_1 - 1) + (n_2 - 1)} \quad (9.12)$$

Eroarea standard a diferenței dintre mediile celor 2 eșantioane, prezentată la numitorul relației 9.11, va fi:

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} \quad (9.13)$$

Nivelul de semnificație $\alpha = 0{,}05$ va fi împărțit în mod egal între cele două cozi laterale $\alpha/2 = 0{,}025$ asociat cu valorile foarte mici și $\alpha/2 = 0{,}025$ asociat cu valorile foarte mari.

Pentru determinarea valorii critice $t_{n-1,\ 1-\alpha/2}$ putem folosi tabelele distribuției t, sau formula Excel: $= T.INV(0{,}975;\ n_1 + n_2 - 2)$

Regiunea de respingere a testului Student „t" va fi în consecință compusă din 2 părți:

$$t_{calculat} \leq -t_{n_1+n_2-2,\ 1-\alpha/2} \quad \text{și} \quad t_{calculat} \geq t_{n_1+n_2-2,\ 1-\alpha/2}$$

Exemplu: în tabelul 9.5 avem nivelurile colesterolului total (mg/dl) măsurate în cazul a 93 de pacienți, 52 cu hipertensiune primară și 41 normotensivi. Putem concluziona că pacienții cu hipertensiune primară au în medie niveluri de colesterol diferite comparativ cu pacienții normotensivi, în populațiile din care provin cele două eșantioane?

Tabelul 9.5. Nivelurile colesterolului (mg/dl) din cele 2 eșantioane

Nr. crt.	Grup pacienți	Colesterol (mg/dl)
1	Hipertensiune primară	207
2	Hipertensiune primară	172
3	Hipertensiune primară	191
4	Hipertensiune primară	221
...	...	...
90	Normotensiv	200
91	Normotensiv	135
92	Normotensiv	161
93	Normotensiv	203

Scanned with OKEN Scanner

Condiții preliminare: datele colectate reprezintă două eșantioane care conțin nivelurile colesterolului total extrase aleatoriu din 2 populații, unul dintr-o populație de pacienți cu hipertensiune primară, iar celălalt dintr-o populație de pacienți normotensivi. Presupunem faptul că nivelurile colesterolului total în ambele populații sunt normal distribuite, iar cele două varianțe, necunoscute la nivel populațional, le considerăm egale ($\sigma_1^2 = \sigma_2^2$).

Ipotezele statistice ale studiului vor fi:

- Ipoteza nulă H_0: nivelurile medii ale colesterolului în cele 2 populații sunt egale ($\mu_1 = \mu_2$ sau $\mu_1 - \mu_2 = 0$)
- Ipoteza alternativă H_1: nivelurile medii ale colesterolului în cele 2 populații nu sunt egale ($\mu_1 \neq \mu_2$ sau $\mu_1 - \mu_2 \neq 0$)

Atunci când ipoteza nulă H_0 este adevărată, testul statistic urmează o distribuție Student „t" cu $n_1 + n_2 - 2 = 93 - 2 = 91$ grade de libertate.

Nivelul de semnificație $\alpha = 0{,}05$ va fi împărțit în mod egal între cele două cozi laterale $\alpha/2 = 0{,}025$ asociat cu valorile foarte mici și $\alpha/2 = 0{,}025$ asociat cu valorile foarte mari (figura 9.10).

Valoarea critică a testului t se citește din tabelul distribuției Student. Alternativ, se poate utiliza funcția Excel: $= T.INV(0{,}975; 91)$:

$$t_{critic\ bilateral} = t_{91,\ 0.975} = 1{,}986$$

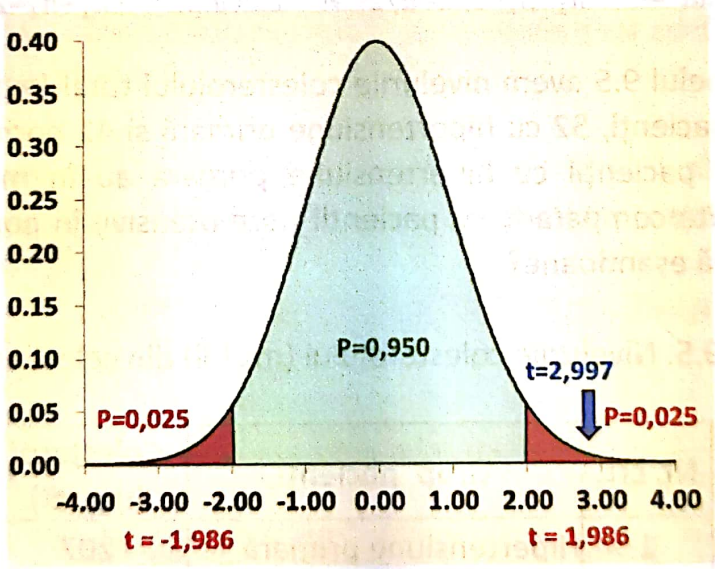


Figura 9.10. Distribuția valorilor testului statistic t bilateral cu 91 de grade de libertate pentru toate perechile de eșantioane extrase

Calculăm indicatorii statistici descriptivi (media și varianța), la nivelul fiecărui eșantion:

$$\bar{x}_{hipertensivi} = \frac{\sum_{i=1}^{52} x_i}{n_1} = \frac{11020}{52} = 211{,}923\ \text{mg/dl}$$

Scanned with OKEN Scanner

$$s^2_{hipertensivi} = \frac{\sum_{i=1}^{52}(x_i - \bar{x})^2}{n_1 - 1} = \frac{37783{,}692}{52 - 1} = 740{,}857 \text{ mg/dl}$$

$$\bar{x}_{normotensivi} = \frac{\sum_{i=1}^{41} x_i}{n_2} = \frac{7921}{41} = 193{,}195 \text{ mg/dl}$$

$$s^2_{normotensivi} = \frac{\sum_{i=1}^{41}(x_i - \bar{x})^2}{n_2 - 1} = \frac{43678{,}439}{41 - 1} = 1091{,}961 \text{ mg/dl}$$

Presupunând faptul că varianțele celor două populații, din care provin eșantioanele de lucru, sunt egale vom determina varianța comună a acestora prin aplicarea relației 9.12:

$$s_p^2 = \frac{s_1^2 \cdot (n_1 - 1) + s_2^2 \cdot (n_2 - 1)}{(n_1 - 1) + (n_2 - 1)} = \frac{740{,}857 \cdot 51 + 1091{,}961 \cdot 40}{51 + 40}$$

Efectuând calculele, obținem: $s_p^2 = 895{,}188$ mg/dl

Testul statistic utilizat pentru a testa ipoteza nulă H_0: $\mu_1 = \mu_2$, conform relației 9.11 este:

$$t_{calculat} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{211{,}923 - 193{,}195}{\sqrt{\frac{895{,}188}{52} + \frac{895{,}188}{41}}} = 2{,}997$$

Vom respinge H_0, dacă valoarea calculată a testului este cuprinsă în regiunea de respingere, pentru nivelul de semnificație $\alpha = 0{,}05$. În baza datelor furnizate de eșantionul nostru avem $t_{calculat} = 2{,}997$, valoare situată în mod evident în regiunea de respingere (zona marcată cu roșu din figura 9.10).

Prin urmare, vom respinge ipoteza nulă H_0 și vom afirma că valoarea calculată a testului statistic calculat t este semnificativă la un nivel $\alpha = 0{,}05$. Rezultatele testului sugerează că există o diferență semnificativă statistic între nivelurile medii ale colesterolului în cele 2 populații (pacienții cu hipertensiune primară și pacienții normotensivi).

Realizând analiza în programul SPSS vom urma următoarele etape:

a) calculăm indicatorii statistici descriptivi, la nivelul fiecărui eșantion:

Group Statistics

	Tip_pacienti	N	Mean	Std. Deviation	Std. Error Mean
Colesterol_total	Hipertensiune primara	52	211.9231	27.21868	3.77455
	Normotensivi	41	193.1951	33.04483	5.16074

Scanned with OKEN Scanner

b) realizăm o reprezentare grafică comparativă a celor două valori medii, împreună cu nivelul variabilității specific fiecărui eșantion, evidențiat de deviația standard (SD):

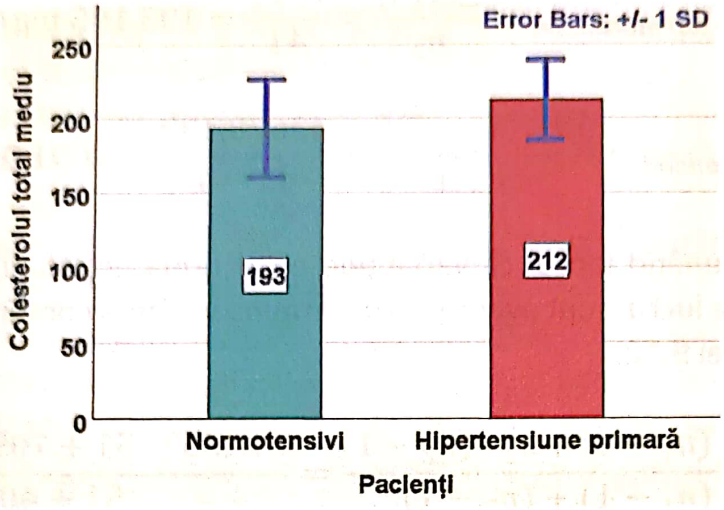


Figura 9.11. Mediile și deviațiile standard ale celor două eșantioane

c) evaluăm normalitatea distribuțiilor populațiilor din care provin eșantioanele de studiu prin utilizarea *testelor statistice de normalitate*:

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Tip_pacienti	Statistic	df	Sig.	Statistic	df	Sig.
Colesterol_total	Normotensivi	.130	41	.080	.950	41	.070
	Hipertensiune primara	.130	52	.027	.963	52	.111

a. Lilliefors Significance Correction

d) cercetăm egalitatea variațelor din populațiile generale, din care au fost extrase în mod aleatoriu cele două eșantioane, prin aplicarea testului Levene; acesta verifică ipoteza nulă conform căreia variațele populațiilor din care provin grupurile sunt egale $H_0: \sigma_1^2 = \sigma_2^2$ și este utilizat anterior aplicării testului Student „t", pentru a putea decide între cele două variante ale acestui test:

Levene's Test for Equality of Variances

F	Sig.
.825	.366

Scanned with OKEN Scanner

În situația în care $p \geq 0{,}05$, nu respingem ipoteza nulă H_0 și considerăm că variațele populațiilor din care provin eșantioanele ar putea fi egale (în cazul nostru $p = 0{,}366 > 0{,}05$).

e) Calculăm valoarea testului statistic Student „t", utilizat pentru a verifica ipoteza nulă H_0: $\mu_1 = \mu_2$. Vom alege în consecință, prima valoare din tabel $t = 2{,}997$, corespunzătoare situației în care variațele celor două populații sunt considerate egale *„Equal variances assumed"*. Programul va evidenția și în acest caz intervalul de încredere 95% pentru diferența mediilor celor două populații.

Independent Samples Test

		t-test for Equality of Means						
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference Lower	Upper
Colesterol_total	Equal variances assumed	2.997	91	.004	18.72795	6.24892	6.31524	31.14067
	Equal variances not assumed	2.929	76.967	.004	18.72795	6.39378	5.99622	31.45969

Întrucât, în baza testului Levene calculat anterior am considerat că variațele din cele două populații studiate ar putea fi egale, vom aplica prima variantă a testului t (presupunând $\sigma_1^2 = \sigma_2^2$).

Intervalul de încredere 95% determinat pentru diferența mediilor celor două populații: (6,315; 31,141) nu include valoarea zero (exclude deci posibilitatea egalității mediilor) și prin urmare putem considera că media colesterolului în populația pacienților cu hipertensiune primară este semnificativ diferită de media colesterolului pacienților normotensivi. De asemenea, probabilitatea obținerii unei valori a testului Student cel puțin la fel „de extremă" ca valoarea calculată la nivelul eșantioanelor noastre: $t_{calculat} = 2{,}997$ este $p = 0{,}004$, aceasta fiind mult mai redusă în comparație cu nivelul de semnificație $\alpha = 0{,}05$, fixat la începutul studiului. Prin urmare vom respinge ipoteza nulă H_0, a egalității valorilor medii din cele două populații, și vom concluziona că diferența dintre acestea este semnificativă statistic.

C. eșantioanele sunt independente și provin din 2 populații normal distribuite având variațele (σ_1^2, σ_2^2) necunoscute și inegale

Testul statistic utilizat pentru a testa ipoteza nulă H_0: $\mu_1 = \mu_2$ sau $\mu_1 - \mu_2 = 0$ (H_1: $\mu_1 \neq \mu_2$ sau $\mu_1 - \mu_2 \neq 0$) este:

$$t_{calculat} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \qquad (9.14)$$

Scanned with OKEN Scanner

care, atunci când H_0 este adevărată, urmează aproximativ o *distribuție Welch* t_r (o variantă adaptată a distribuției Student), unde r reprezintă valoarea ajustată a gradelor de libertate, care se determină cu următoarea formulă:

$$r = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} \tag{9.15}$$

Întrucât valoarea ajustată a gradelor de libertate r este în cele mai multe cazuri un număr real, putem utiliza în analiză doar partea întreagă a acestuia.

Nivelul de semnificație $\alpha = 0{,}05$ va fi împărțit în mod egal între cele două cozi laterale $\alpha/2 = 0{,}025$ asociat cu valorile foarte mici și $\alpha/2 = 0{,}025$ asociat cu valorile foarte mari.

Pentru determinarea valorii critice $t_{r,1-\alpha/2}$ putem folosi tabelele distribuției t, sau formula Excel: $= T.INV(0{,}975; r)$.

Regiunea de respingere a testului t_r va fi compusă din 2 părți:

$$t_{calculat} \leq -t_{r,1-\alpha/2} \quad \text{și} \quad t_{calculat} \geq t_{r,1-\alpha/2}$$

Exemplu: în tabelul 9.6 au fost măsurate diametrele (μm) pentru 40 de limfocite și 50 de celule tumorale obținute în urma biopsiei tisulare la pacienții cu melanom. Putem concluziona faptul că, în medie, limfocitele și celulele tumorale diferă ca mărime în populațiile din care provin cele două eșantioane?

Tabelul 9.6. Diametrele limfocitelor și celulelor tumorale (μm) din cele două eșantioane

Nr. crt.	Tip celulă	Diametru (μm)
1	Limfocite	9,00
2	Limfocite	6,30
3	Limfocite	8,60
4	Limfocite	7,40
...	...	...
87	Celule tumorale	18,70
88	Celule tumorale	15,20
89	Celule tumorale	20,70
90	Celule tumorale	21,50

Scanned with OKEN Scanner

Condiții preliminare: datele colectate reprezintă două eșantioane care conțin nivelurile diametrelor celulelor (μm) extrase aleatoriu din 2 populații, una de limfocite și a doua de celule tumorale, recoltate de la pacienți cu melanom. Presupunem faptul că diametrul celulelor din fiecare populație urmează o distribuție normală, iar cele două varianțe necunoscute la nivel populațional le considerăm inegale.

Ipotezele statistice ale studiului vor fi:

- Ipoteza nulă H_0: diametrele medii ale limfocitelor și celulelor tumorale în cele 2 populații sunt egale ($\mu_1 = \mu_2$ sau $\mu_1 - \mu_2 = 0$)
- Ipoteza alternativă H_1: diametrele medii ale limfocitelor și celulelor tumorale în cele 2 populații nu sunt egale ($\mu_1 \neq \mu_2$ sau $\mu_1 - \mu_2 \neq 0$)

Calculăm indicatorii statistici descriptivi (media și varianța), la nivelul fiecărui eșantion:

$$\bar{x}_{celule\ tumorale} = \frac{\sum_{i=1}^{50} x_i}{n_2} = \frac{896}{50} = 17{,}920\ \mu m$$

$$s^2_{celule\ tumorale} = \frac{\sum_{i=1}^{50}(x_i - \bar{x})^2}{n_2 - 1} = \frac{431{,}82}{50 - 1} = 8{,}813\ \mu m$$

$$\bar{x}_{limfocite} = \frac{\sum_{i=1}^{40} x_i}{n_1} = \frac{278}{40} = 6{,}950\ \mu m$$

$$s^2_{limfocite} = \frac{\sum_{i=1}^{40}(x_i - \bar{x})^2}{n_1 - 1} = \frac{99{,}320}{40 - 1} = 2{,}547\ \mu m$$

Atunci când ipoteza nulă este adevărată, testul statistic urmează o distribuție t_r Welch cu r grade de libertate, unde r se determină folosind relația 9.15:

$$r = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}} = \frac{\left(\frac{8{,}813}{50} + \frac{2{,}547}{40}\right)^2}{\frac{\left(\frac{8{,}813}{50}\right)^2}{50 - 1} + \frac{\left(\frac{2{,}547}{40}\right)^2}{40 - 1}} = 78{,}005$$

Rotunjind, obținem $r = 78$ grade de libertate.

Nivelul de semnificație $\alpha = 0{,}05$ va fi împărțit în mod egal între cele două cozi laterale.

Scanned with OKEN Scanner

Valoarea critică a testului Student: $t_{critic\ bilateral} = t_{78,\ 0.975} = 1{,}991$
Valoarea critică se citește din tabelul distribuției Student sau se poate obține utilizând următoarea funcție din programul Excel: $= T.INV(0{,}975; 78)$

Testul statistic utilizat pentru a testa ipoteza nulă H_0: $\mu_1 = \mu_2$, conform relației 9.14 este:

$$t_{calculat} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{17{,}920 - 6{,}950}{\sqrt{\frac{8{,}813}{50} + \frac{2{,}547}{40}}} = 22{,}396$$

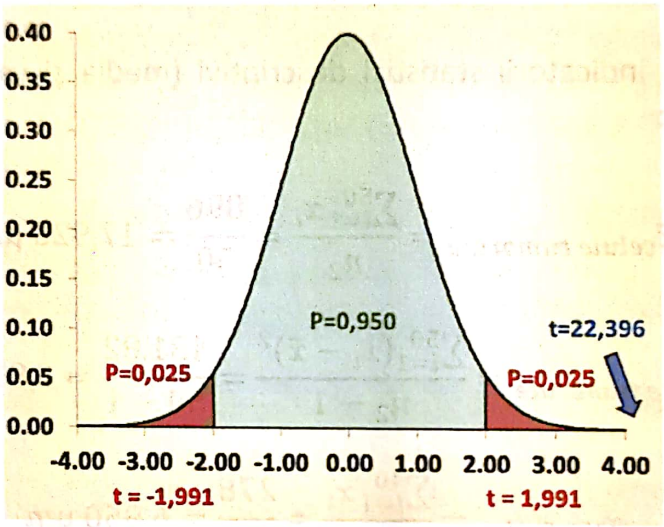


Figura 9.12. Distribuția valorilor testului statistic t bilateral cu 78 de grade de libertate pentru toate perechile de eșantioane extrase

Vom respinge H_0, dacă valoarea calculată a testului este cuprinsă în regiunea de respingere, pentru nivelul de semnificație $\alpha = 0{,}05$. În baza datelor furnizate de eșantionul nostru avem $t_{calculat} = 22{,}396$, valoare situată în mod evident în regiunea de respingere (zona marcată cu roșu din figura 9.12).

În consecință, vom respinge ipoteza nulă H_0 și vom considera că valoarea calculată a testului statistic calculat t este semnificativă la un nivel $\alpha = 0{,}05$. Rezultatele obținute sugerează faptul că există o diferență semnificativă statistic între diametrele medii ale limfocitelor și celulelor tumorale, în cazul pacienților cu melanom.

Analiza în programul SPSS se realizează similar cazului anterior prezentat la punctul B:

a) calculăm indicatorii statistici descriptivi, la nivelul fiecărui eșantion:

Scanned with OKEN Scanner

Group Statistics

	Tip_celula	N	Mean	Std. Deviation	Std. Error Mean
Diametru	Limfocite	40	6.9500	1.59583	.25232
	Celule tumorale	50	17.9200	2.96861	.41983

b) realizăm o reprezentare grafică comparativă a celor două valori medii, împreună cu nivelul variabilității specific fiecărui eșantion

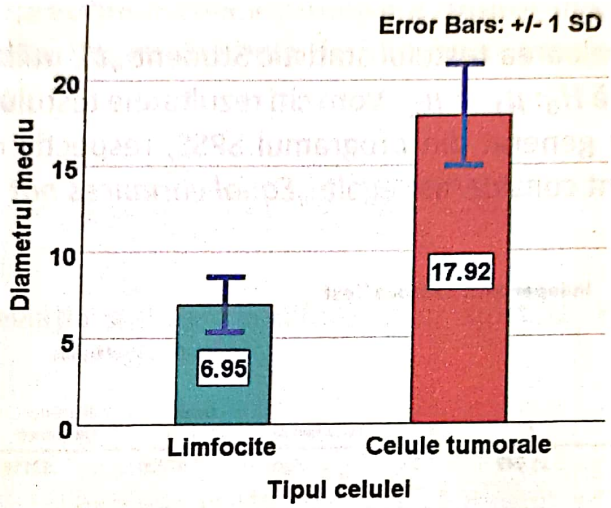


Figura 9.13. Mediile și deviațiile standard ale celor două eșantioane

c) evaluăm normalitatea distribuțiilor populațiilor din care provin eșantioanele de studiu prin utilizarea *testelor statistice de normalitate*:

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Tip_celula	Statistic	df	Sig.	Statistic	df	Sig.
Diametru	Limfocite	.076	40	.200*	.978	40	.605
	Celule tumorale	.066	50	.200*	.987	50	.860

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

d) cercetăm egalitatea varianțelor din populațiile generale, din care au fost extrase în mod aleatoriu cele două eșantioane, prin aplicarea testului Levene:

Scanned with OKEN Scanner

Levene's Test for Equality of Variances	
F	Sig.
9.684	.003

Deoarece, în acest caz $p = 0{,}003 < 0{,}05$, vom respinge ipoteza nulă a egalității variațelor populațiilor din care provin grupurile $H_0: \sigma_1^2 = \sigma_2^2$, și în consecință vom utiliza varianta testului t care presupune că variațele nu sunt egale.

e) Calculăm valoarea testului statistic Student „t", utilizat pentru a verifica ipoteza nulă $H_0: \mu_1 = \mu_2$. Vom citi rezultatele testului t de pe linia a doua a tabelului generat din programul SPSS, respectiv cea pentru variațe care nu sunt considerate egale *„Equal variances not assumed"*.

Independent Samples Test

		t-test for Equality of Means						
							95% Confidence Interval of the Difference	
		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Diametru	Equal variances assumed	21.049	88	.000	10.97000	.52116	9.93431	12.00569
	Equal variances not assumed	22.396	78.005	.000	10.97000	.48982	9.99485	11.94515

Probabilitatea obținerii unei valori a testului Student cel puțin la fel „de extremă" cum este valoarea calculată la nivelul eșantioanelor noastre: $t_{calculat} = 22{,}396$ este $p = 0{,}000$ (valoare foarte apropiată de zero), aceasta fiind mult mai redusă în comparație cu nivelul de semnificație $\alpha = 0{,}05$, fixat la începutul studiului. Prin urmare vom respinge ipoteza nulă H_0, a egalității valorilor medii din cele două populații, și vom concluziona că diferența dintre diametrele medii ale limfocitelor și celulelor tumorale este semnificativă statistic la nivel populațional.

D. eșantioanele sunt dependente (datele sunt perechi) și provin dintr-o populație normal distribuită

În locul studierii valorilor individuale, vom utiliza în analiză diferența între fiecare pereche de observații a variabilei de interes. Cele n diferențe calculate la nivel de eșantion între cele n perechi de valori reprezintă un eșantion aleatoriu extras dintr-o populație normal distribuită a diferențelor $d_i = x_{i1} - x_{i0}$. Observațiile pereche pot fi reprezentate de aceeași subiecți analizați înainte (x_{i0}) și după aplicarea unui anumit tip de tratament (x_{i1}), în scopul evaluării eficienței acestuia. Un alt exemplu ar putea fi reprezentat de aplicarea a două metode de tratament diferite aceluiași grup de pacienți

Scanned with OKEN Scanner

Testul statistic utilizat pentru a testa ipoteza nulă $H_0: \mu_d = 0$ versus ipoteza alternativă $H_1: \mu_d \neq 0$ este:

$$t_{calculat} = \frac{\bar{d} - \mu_{d_0}}{\frac{s_{\bar{d}}}{\sqrt{n}}} = \frac{\bar{d}}{\frac{s_{\bar{d}}}{\sqrt{n}}} \tag{9.16}$$

care, atunci când H_0 este adevărată, urmează o *distribuție Student* cu $n - 1$ grade de libertate.

În relația 9.16, $\bar{d}$ reprezintă media aritmetică a tuturor diferențelor $d_i = x_{i1} - x_{i0}$ calculate la nivelul fiecărei perechi de observații a variabilei de interes din cadrul eșantionului:

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{\sum_{i=1}^{n}(x_{i1} - x_{i0})}{n} \tag{9.17}$$

iar $s_{\bar{d}}$ reprezintă deviația standard a diferențelor ajustată, calculată la nivel de eșantion:

$$s_{\bar{d}} = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}} \tag{9.18}$$

Nivelul de semnificație al testului $\alpha = 0{,}05$ va fi împărțit în mod egal între cele două cozi laterale $\alpha/2 = 0{,}025$ asociat cu valorile foarte mici și $\alpha/2 = 0{,}025$ asociat cu valorile foarte mari.

Pentru determinarea valorii critice $t_{n-1,\ 1-\alpha/2}$ putem folosi tabelele distribuției Student, sau putem genera direct valoarea din programul Excel, prin aplicarea formulei: $= T.INV(0{,}975; n-1)$.

Regiunea de respingere a testului Student „t" va fi în consecință compusă din 2 părți:

$$t_{calculat} \leq -t_{n-1,\ 1-\alpha/2} \quad \text{și} \quad t_{calculat} \geq t_{n-1,\ 1-\alpha/2}$$

Exemplu: în cadrul unui studiu retrospectiv au fost colectate înregistrările privitoare la un lot de 30 de pacienți care au fost tratați pentru hipercolesterolemie. În tabelul 9.7, sunt prezentate înregistrările valorilor colesterolului total ale pacienților, anterior și după șase săptămâni de la administrarea medicamentului. Avem suficiente dovezi, în cazul unui nivel de semnificație $\alpha = 0{,}05$, pentru a concluziona faptul că administrarea medicamentului va determina o scădere a nivelului total al colesterolului într-o populație similară, formată din pacienți cu hipercolesterolemie?

Scanned with OKEN Scanner

Tabelul 9.7. Nivelurile colesterolului total (mg/dl) ale celor 30 de pacienți înainte și după administrarea medicamentului

Nr. crt.	Înainte	După	Diferențe (d_i)
1	203	160	-43
2	295	171	-124
3	301	182	-119
4	259	145	-114
5	314	201	-113
...	...	...	...
26	266	207	-59
27	321	248	-73
28	233	195	-38
29	224	147	-77
30	274	135	-139

Condiții preliminare: datele sunt colectate în perechi logice: aceeași unitate statistică (pacient) este măsurată de două ori, de exemplu înainte și după o intervenție („înainte de tratament" și „după tratament"). Diferențele dintre perechi trebuie să fie independente una de cealaltă, adică diferența în cazul pacientului 1 nu trebuie să fie influențată de diferența înregistrată pentru pacientul 2 etc. Distribuția diferențelor dintre valorile fiecărei perechi $d_i = x_{i1} - x_{i0}$ trebuie să fie aproximativ normală (nu este necesar ca valorile înregistrate înainte și după tratamentul medicamentos să fie fiecare normal distribuite).

Ipotezele statistice ale studiului vor fi:

- Ipoteza nulă H_0: nu există o diferență semnificativă statistic între nivelul mediu al colesterolului total măsurat înainte și după administrarea tratamentului în populația generală ($\mu_d = 0$)
- Ipoteza alternativă H_1: există o diferență semnificativă statistic între nivelul mediu al colesterolului total măsurat înainte și după administrarea tratamentului în populația generală ($\mu_d \neq 0$)

Atunci când ipoteza nulă H_0 este adevărată, testul statistic urmează o distribuție Student „t" cu $n - 1 = 30 - 1 = 29$ grade de libertate.

Nivelul de semnificație $\alpha = 0{,}05$ va fi împărțit în mod egal între cele două cozi laterale $\alpha/2 = 0{,}025$ asociat cu valorile foarte mici și $\alpha/2 = 0{,}025$ asociat cu valorile foarte mari (figura 9.14).

Valoarea critică a testului t se citește din tabelul distribuției Student. Alternativ, se poate utiliza funcția Excel: $= T.INV(0{,}975; 29)$:

$$t_{critic\ bilateral} = t_{29,\ 0.975} = 1{,}968$$

Scanned with OKEN Scanner

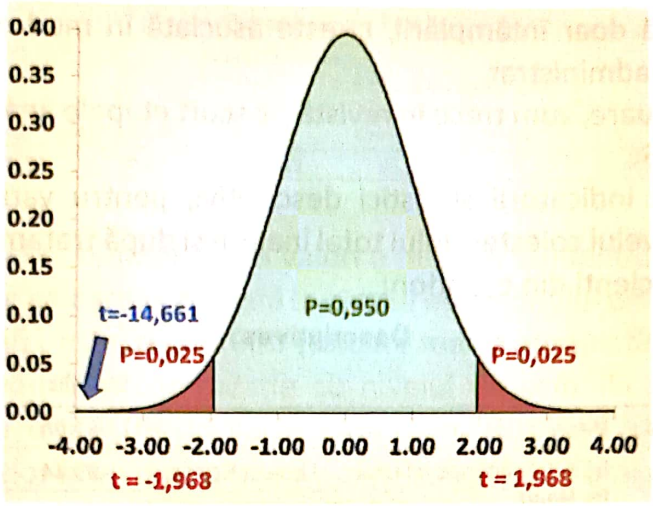


Figura 9.14. Distribuția valorilor testului statistic t bilateral cu 29 de grade de libertate pentru toate perechile de eșantioane extrase

Calculăm media și deviația standard ajustată a celor 30 de diferențe din eșantion (relațiile 9.17 și 9.18):

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} = \frac{-2460}{30} = -82\ mg/dl$$

$$s_{\bar{d}} = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}} = \sqrt{\frac{27216}{30-1}} = \sqrt{938{,}483} = 30{,}635\ mg/dl$$

Testul statistic utilizat pentru a testa ipoteza nulă $H_0: \mu_d = 0$ versus ipoteza alternativă $H_1: \mu_d \neq 0$ este:

$$t_{calculat} = \frac{\bar{d}}{\frac{s_{\bar{d}}}{\sqrt{n}}} = \frac{-82}{\frac{30{,}635}{\sqrt{30}}} = -14{,}661$$

Având în vedere faptul că $t_{calculat} < -t_{critic\ bilateral}$, rezultatul obținut este semnificativ statistic (valoarea obținută a testului în baza datelor din eșantion se situează în extremitatea stângă a regiunii de respingere, evidențiată în figura 9.12). Prin urmare, eșantionul nostru furnizează suficiente dovezi în sprijinul respingerii ipotezei nule $H_0: \mu_d = 0$, și astfel considerăm că administrarea medicamentului poate determina o reducere a nivelului total al colesterolului la nivel populațional. În concluzie, diferența medie calculată de $-82\ mg/dl$, dintre valorile colesterolului total înainte și după tratament, nu

Scanned with OKEN Scanner

poate fi atribuită doar întâmplării, ci este asociată în mod evident cu efectul medicamentului administrat.

În continuare, vom trece în revistă pe scurt etapele analizei prin utilizarea programului SPSS:

a) calculăm indicatorii statistici descriptivi, pentru variabila „diferențele dintre nivelul colesterolului total înainte și după tratament" în cazul celor 30 de pacienți din eșantion:

Descriptives

			Statistic	Std. Error
Diferente	Mean		-82.00	5.593
	95% Confidence Interval for Mean	Lower Bound	-93.44	
		Upper Bound	-70.56	
	5% Trimmed Mean		-81.98	
	Median		-76.50	
	Variance		938.483	
	Std. Deviation		30.635	
	Minimum		-139	
	Maximum		-21	
	Range		118	
	Interquartile Range		45	
	Skewness		-.073	.427
	Kurtosis		-.705	.833

b) evaluăm normalitatea distribuției populației diferențelor valorilor colesterolului total înainte și după tratament, din care provine eșantionul celor 30 de diferențe studiat, prin utilizarea *testelor statistice de normalitate*:

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Diferente	.106	30	.200*	.977	30	.733

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

c) Calculăm valoarea testului statistic Student „t", utilizat pentru a verifica ipoteza nulă H_0: $\mu_d = 0$. Așa cum am observat și în situațiile anterioare, programul calculează si intervalul de încredere 95% pentru diferența medie a valorilor colesterolului total înainte și după tratament.

Scanned with OKEN Scanner

Paired Samples Test

		Paired Differences							
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig. (2-tailed)
Pair 1	Dupa - Inainte	-82.000	30.635	5.593	-93.439	-70.561	-14.661	29	.000

Probabilitatea obținerii unei valori a testului Student cel puțin la fel „de extremă" cum este valoarea calculată la nivelul eșantionului diferențelor noastre: $t_{calculat} = -14{,}661$ este $p = 0{,}000$ (valoare foarte apropiată de zero), aceasta fiind mult mai redusă în comparație cu nivelul de semnificație fixat anterior derulării studiului $\alpha = 0{,}05$. În consecință, vom respinge ipoteza nulă H_0, a absenței diferențelor între nivelul mediu al colesterolului total măsurat înainte și după administrarea tratamentului și vom concluziona faptul că scăderea medie a valorii colesterolului total de $82\,mg/dl$ este datorată administrării medicamentului. În concluzie, apreciem cu un nivel de confidență de 95%, că intervalul $(-93{,}439;\ -70{,}561)$ va incluse diferența medie reală existentă în populația generală, dintre valorile colesterolului total măsurat înainte și după tratament.

9.5. TESTAREA IPOTEZELOR PENTRU DIFERENȚA PROPORȚIILOR A DOUĂ POPULAȚII - TESTUL „z"

În domeniul medical, luarea deciziilor bazate pe dovezi este esențială pentru evaluarea eficienței diferitelor tratamente, diagnostice sau intervenții chirurgicale. În multe situații practice, se dorește realizarea unei comparații între două proporții – de exemplu, proporția pacienților care s-au vindecat în urma a două tratamente diferite sau proporția persoanelor care dezvoltă reacții adverse în cadrul a două regimuri terapeutice. În astfel de situații, se utilizează în practică testul „z", care realizează o evaluare a diferenței proporțiilor observate la nivelul eșantioanelor și în final oferă suport statistic pentru luarea unei decizii privind populațiile din care au fost extrase în mod aleatoriu aceste eșantioane (diferența este semnificativă statistic sau poate fi atribuită șansei).

Pentru ca testul „z" în cazul diferenței proporțiilor a două populații să furnizeze rezultate valide, eșantioanele studiate trebuie să fie independente și să provină din populații aproximativ normal distribuite, sau cel puțin volumul lor să fie suficient de mare încât să fie aplicabilă Teorema Limită Centrală ($n_1 \cdot p_1 \geq 5,\ n_1 \cdot (1-p_1) \geq 5, n_2 \cdot p_2 \geq 5,\ n_2 \cdot (1-p_2) \geq 5$).

Distribuția de sondaj a diferenței proporțiilor celor două eșantioane $(p_1 - p_2)$, ca estimator punctual al diferenței proporțiilor celor 2 populații $(\pi_1 - \pi_2)$, va fi aproximativ normal distribuită având media $(\pi_1 - \pi_2)$.

Scanned with OKEN Scanner

În baza ipotezei nule $H_0: \pi_1 = \pi_2$ considerăm faptul că proporțiile celor două populații sunt egale (sau proporțiile eșantioanelor p_1 și p_2 estimează aceeași proporție comună $\bar{p}$). Proporția comună se determină după relația:

$$p_1 = p_2 = \bar{p} = \frac{x_1 + x_2}{n_1 + n_2} \tag{9.19}$$

în care x_1 și x_2 reprezintă numărul de unități statistice din fiecare eșantion care prezintă caracteristica de interes, iar n_1 și n_2 volumele celor 2 eșantioane.

Prin urmare, eroarea standard a diferenței dintre proporțiile celor 2 eșantioane va fi:

$$\sigma_{p_1-p_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{p_1 \cdot (1-p_1)}{n_1} + \frac{p_2 \cdot (1-p_2)}{n_2}}$$

Înlocuind proporțiile specifice fiecărui eșantion (p_1 și p_2) cu proporția comună estimată $\bar{p}$, calculată conform relației 9.19, obținem:

$$\sigma_{p_1-p_2} = \sqrt{\frac{\bar{p} \cdot (1-\bar{p})}{n_1} + \frac{\bar{p} \cdot (1-\bar{p})}{n_2}} \tag{9.20}$$

Testul statistic utilizat pentru a testa ipoteza nulă $H_0: \pi_1 = \pi_2$ sau $\pi_1 - \pi_2 = 0$ împotriva ipotezei alternative $H_1: \pi_1 \neq \pi_2$ sau $\pi_1 - \pi_2 \neq 0$ este:

$$z_{calculat} = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sigma_{p_1-p_2}} = \frac{p_1 - p_2}{\sqrt{\frac{\bar{p} \cdot (1-\bar{p})}{n_1} + \frac{\bar{p} \cdot (1-\bar{p})}{n_2}}} \tag{9.21}$$

care, atunci când H_0 este adevărată, urmează o distribuție normală standard.

Nivelul de semnificație $\alpha = 0{,}05$ va fi împărțit în mod egal între cele două cozi laterale $\alpha/2 = 0{,}025$ asociat cu valorile foarte mici și $\alpha/2 = 0{,}025$ asociat cu valorile foarte mari.

Valoarea critică $z_{1-\alpha/2} = 1{,}96$ reprezintă abscisa corespunzătoare probabilității cumulate $P = 1 - \alpha/2 = 1 - 0{,}05/2 = 1 - 0{,}025 = 0{,}975$ se poate citi din tabelele distribuției normale standard sau poate fi obținută utilizând formula Excel: $= NORM.S.INV(0{,}975)$

Regiunea de respingere a testului z va fi în compusă din 2 părți:

$$z_{calculat} \leq -z_{1-\alpha/2} \quad \text{și} \quad z_{calculat} \geq z_{1-\alpha/2}$$

Scanned with OKEN Scanner

Exemplu: Un medicament antihistaminic a fost testat pentru a determina dacă există o diferență între viteza de reacție a pacienților adulți studiați. Dintr-un eșantion extras aleatoriu $n_1 = 200$ de adulți de sex masculin cărora li s-a administrat medicamentul, $x_1 = 20$ mai aveau simptome la 30 de minute după administrare. Dintr-un alt eșantion obținut aleatoriu $n_2 = 200$ de adulți de sex feminin cărora li s-a administrat același medicament, $x_2 = 12$ mai prezentau simptome la 30 de minute după administrare. Testați dacă există o diferență semnificativă statistic între procentele pacienților de sex masculin ($p_1 = 20/200 = 0{,}10$) și feminin ($p_2 = 12/200 = 0{,}06$) care nu au reacționat la administrarea medicamentului (prezintă in continuare simptome după 30 de minute de la administrare), pentru un nivel de semnificație $\alpha = 0{,}05$.

Condiții preliminare:

- cele 2 eșantioane, extrase în mod aleatoriu din două populații de adulți cu alergie (M/ F), sunt independente;
- deoarece volumele celor două eșantioane sunt suficient de mari $n_1 = n_2 = 200$ persoane, condițiile $n_1 \cdot p_1 \geq 5$, $n_1 \cdot (1 - p_1) \geq 5, n_2 \cdot p_2 \geq 5$, $n_2 \cdot (1 - p_2) \geq 5$ sunt îndeplinite:

$n_1 \cdot p_1 = 200 \cdot 0{,}10 = 20 \geq 5$, $n_1 \cdot (1 - p_1) = 200 \cdot 0{,}90 = 180 \geq 5$,
$n_2 \cdot p_2 = 200 \cdot 0{,}06 = 12 \geq 5$, $n_2 \cdot (1 - p_2) = 200 \cdot 0{,}94 = 188 \geq 5$

și, în baza Teoremei limite centrale, rezultă faptul că: distribuția de sondaj a diferenței proporțiilor celor două eșantioane $(p_1 - p_2)$ va fi aproximativ normal distribuită, având media $(\pi_1 - \pi_2)$.

Ipotezele statistice ale studiului vor fi:

- Ipoteza nulă H_0: nu există o diferență semnificativă statistic între procentele pacienților de sex feminin și masculin care nu au reacționat la administrarea medicamentului în populația generală:
$$\pi_1 = \pi_2 \text{ sau } \pi_1 - \pi_2 = 0$$
- Ipoteza alternativă H_1: există o diferență semnificativă statistic între procentele pacienților de sex feminin și masculin care nu au reacționat la administrarea medicamentului în populația generală:
$$\pi_1 \neq \pi_2 \text{ sau } \pi_1 - \pi_2 \neq 0$$

Atunci când ipoteza nulă H_0 este adevărată, testul statistic urmează o distribuție normală standard (z).

Nivelul de semnificație $\alpha = 0{,}05$ va fi împărțit în mod egal între cele două cozi laterale $\alpha/2 = 0{,}025$ asociat cu valorile foarte mici și $\alpha/2 = 0{,}025$ asociat cu valorile foarte mari (figura 9.15).

Valoarea critică a testului z se citește din tabelele distribuției normale standard sau poate fi obținută utilizând formula Excel: $= NORM.S.INV(0{,}975)$:

$$z_{critic\ bilateral} = z_{1-0{,}05/2} = z_{0{,}975} = 1{,}96$$

Scanned with OKEN Scanner

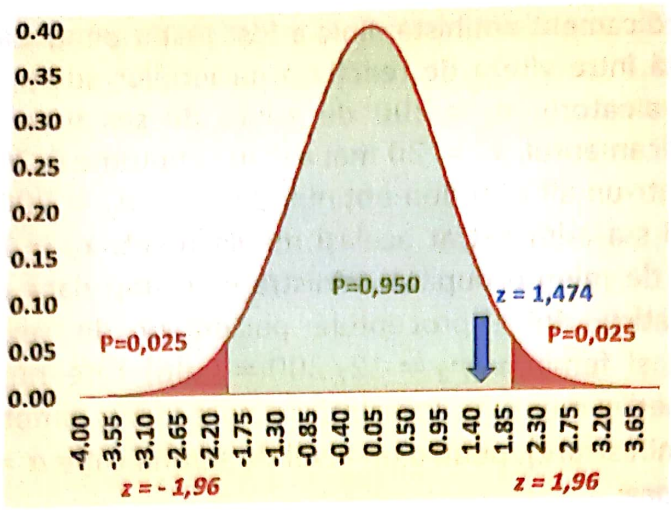


Figura 9.15. Distribuția valorilor testului statistic z bilateral pentru toate perechile de eșantioane extrase

Calculăm proporția comună (relația 9.19):

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{20 + 12}{200 + 200} = 0{,}08$$

Eroarea standard a diferenței dintre proporțiile celor 2 eșantioane va fi (relația 9.20):

$$\sigma_{p_1 - p_2} = \sqrt{\frac{\bar{p} \cdot (1 - \bar{p})}{n_1} + \frac{\bar{p} \cdot (1 - \bar{p})}{n_2}} = \sqrt{\frac{0{,}08(1 - 0{,}08)}{200} + \frac{0{,}08(1 - 0{,}08)}{200}}$$

Efectuând calculele obținem: $\sigma_{p_1 - p_2} = 0{,}027$

Testul statistic utilizat pentru a testa ipoteza nulă a absenței diferențelor între procentele pacienților de sex feminin și masculin care nu au reacționat la administrarea medicamentului este (formula 9.21):

$$z_{calculat} = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sigma_{p_1 - p_2}} = \frac{p_1 - p_2}{\sigma_{p_1 - p_2}} = \frac{0{,}10 - 0{,}06}{0{,}027} = 1{,}474$$

Deoarece valoarea calculată a testului z nu se situează în zona de respingere (figura 9.13):

$$z_{calculat} = 1{,}474 < z_{critic\ bilateral} = 1{,}96$$

considerăm faptul că nu avem suficiente dovezi pentru a concluziona că există o diferență semnificativă statistic între procentele pacienților de sex feminin și

Scanned with OKEN Scanner

masculin care nu au reacționat la administrarea celor 2 medicamente, pentru un nivel de semnificație de $\alpha = 5\%$.

Se citește din tabelul distribuției normale standard probabilitatea cumulată aferentă valorii calculate a testului $z_{calculat} = 1{,}474$, care este $p = 0{,}930$ (în Excel putem utiliza funcția $= NORM.S.DIST(1{,}474; TRUE)$, care returnează același rezultat). Astfel, când H_0 este adevărată, probabilitatea de a obține o valoare a lui z cel puțin la fel de mare ca $1{,}474$ este: $1 - p = 1 - 0{,}930 = 0{,}07$ și, fiind un test bilateral, probabilitatea de a observa o valoare a lui z cel puțin la fel de scăzută ca -1,474 este tot 0,07 (curba normală este simetrică).

În consecință, în cazul exemplului nostru, avem $p = 0{,}14$ (obținut prin însumarea celor două probabilități $0{,}07 + 0{,}07$). Cu alte cuvinte, probabilitatea de a obține o valoare cel puțin la fel de „extremă" cum este cea calculată la nivelul eșantioanelor noastre, respectiv $1{,}474$, în ambele direcții, când ipoteza nulă este adevărată este 0,14 (sau 14%).

9.6. TESTAREA IPOTEZELOR PENTRU RAPORTUL VARIANȚELOR A DOUĂ POPULAȚII - TESTUL „*F*"

Așa cum am văzut anterior, în evaluarea condițiilor preliminare ale testului Student „*t*" pentru diferența dintre mediile a două populații, trebuie să stabilim dacă variațele populațiilor pot fi considerate egale și în consecință să aplicăm testul t corespunzător. Singurele informații de care dispunem însă privind variațele populațiilor generale sunt cele furnizate de cele două eșantioane de studiu, extrase în mod aleatoriu din acestea.

În cadrul analizei concrete, dorim să apreciem dacă diferența, care în mod inevitabil va exista între variațele celor două eșantioane studiate, ne indică o diferență reală între variațele populațiilor de referință sau aceasta este doar rezultatul întâmplării, variațele acestor populații fiind de fapt egale.

Deciziile privind compararea variațelor a două populații au la bază testarea statistică a raportului acestora, pornind de la ipoteză nulă a egalității lor ($\sigma_1^2 = \sigma_2^2$). Atunci când testăm ipoteza egalității variațelor a două populații, investigăm ipoteza că raportul lor este egal cu 1, prin aplicarea testului Fisher „*F*".

În situația în care eșantioanele sunt independente și provin din 2 populații aproximativ normal distribuite, raportul $(s_1^2/\sigma_1^2)/(s_2^2/\sigma_2^2)$ urmează o distribuție Fisher cu $n_1 - 1$ grade de libertate la numărător și $n_2 - 1$ grade de libertate la numitor. Testul F este foarte sensibil la abaterile de la normalitate ale celor două populații și, din acest motiv, atunci când presupunerile de normalitate nu sunt rezonabile nu se recomandă utilizarea lui în practică (se pot utiliza în acest caz teste mai robuste, cum ar fi: Levene sau Bonett, care se bazează doar pe ipoteza că variabilele sunt de tip cantitativ).

Scanned with OKEN Scanner

Testul statistic utilizat pentru a testa ipoteza nulă $H_0: \sigma_1^2 = \sigma_2^2$ împotriva ipotezei alternative $H_1: \sigma_1^2 \neq \sigma_2^2$ este:

$$F_{calculat} = \frac{s_1^2 / \sigma_1^2}{s_2^2 / \sigma_2^2} \tag{9.22}$$

Deoarece lucrăm în condițiile unei ipoteze nule adevărate, respectiv $\sigma_1^2 = \sigma_2^2$, cele două varianțe de mai sus se reduc, și relația 9.22 devine:

$$F_{calculat} = \frac{s_1^2}{s_2^2} \tag{9.23}$$

care, atunci când H_0 este adevărată, urmează o distribuție Fisher cu $n_1 - 1$ grade de libertate la numărător și $n_2 - 1$ grade de libertate la numitor.

Pentru a beneficia de avantajele unui test F unilateral dreapta, varianța cea mai mare va fi poziționată la numărătorul raportului:

$$H_0: \sigma_1^2 \leq \sigma_2^2 \quad \text{și} \quad H_1: \sigma_1^2 > \sigma_2^2$$

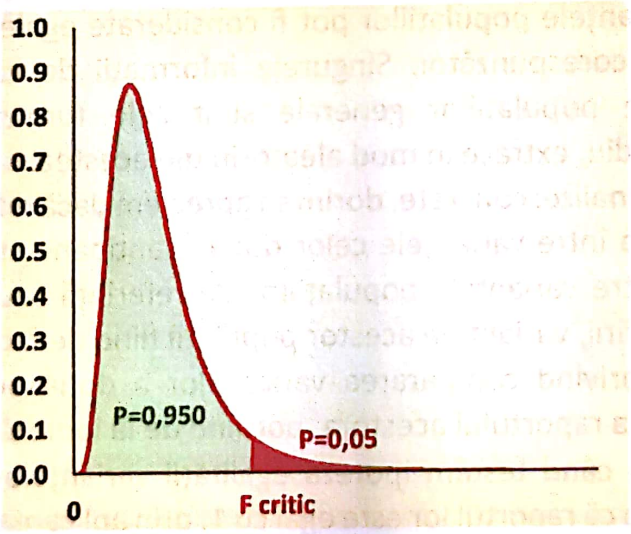


Figura 9.16. Distribuția Fisher cu $n_1 - 1$ și $n_2 - 1$ grade de libertate

Pentru un nivel de semnificație $\alpha = 0{,}05$, se determină valoarea critică $F_{critic} = F_{\alpha, n_1-1, n_2-1}$ din tabelele distribuției Fisher sau aplicând formula Excel $= F.INV.RT(\alpha, n_1 - 1, n_2 - 1)$, corespunzătoare probabilității $P(F > F_{\alpha, n_1-1, n_2-1})$. La dreapta valorii critice a lui F se situează 0,05 sau 5% din aria situată sub curba distribuției Fisher (figura 9.16). De exemplu, în cazul unei distribuții Fisher cu 15 grade de libertate la numărător și 15 la numitor, pentru

Scanned with OKEN Scanner

determinarea valorii critice a testului F vom putea utiliza formula Excel: $= F.INV(0{,}95; 15; 15)$, care va returna valoarea 2,403.

În cazul testului unilateral dreapta, probabilitatea erorii de tip I $\alpha = 0{,}05$ este concentrată în extremitatea dreaptă a distribuției Fisher (figura 9.16).

Dacă $F_{calc} > F_{critic}$, raportul varianțelor se situează în zona de respingere, în extremitatea dreaptă a distribuției (figura 9.14).

Exemplu: revenim la datele privind nivelurile colesterolului total (mg/dl) măsurate în cazul a 93 de pacienți, 52 cu hipertensiune primară și 41 normotensivi. Suntem interesați să testăm dacă presupunerea de egalitate a varianțelor celor două populații din care provin eșantioanele de studiu poate fi luată în considerare pentru a putea utiliza în consecință testul „t" corespunzător.

Condiții preliminare: datele colectate reprezintă două eșantioane care conțin nivelurile colesterolului total extrase aleatoriu din 2 populații, unul dintr-o populație de persoane cu hipertensiune primară, iar celălalt dintr-o populație de persoane normotensive. Presupunem faptul că nivelurile colesterolului total în ambele populații sunt aproximativ normal distribuite.

Ipotezele statistice ale studiului vor fi:

- Ipoteza nulă H_0: $\sigma_1^2 \leq \sigma_2^2$
- Ipoteza alternativă H_1: $\sigma_1^2 > \sigma_2^2$

Testul statistic utilizat pentru a testa ipoteza nulă H_0 (relația (9.23):

$$F_{calculat} = \frac{s_1^2}{s_2^2} = \frac{1091{,}961}{740{,}857} = 1{,}474$$

care, atunci când H_0 este adevărată urmează, o distribuție Fisher cu $n_1 - 1 = 41 - 1 = 40$ grade de libertate la numărător și $n_2 - 1 = 52 - 1 = 51$ grade de libertate la numitor (figura 9.17).

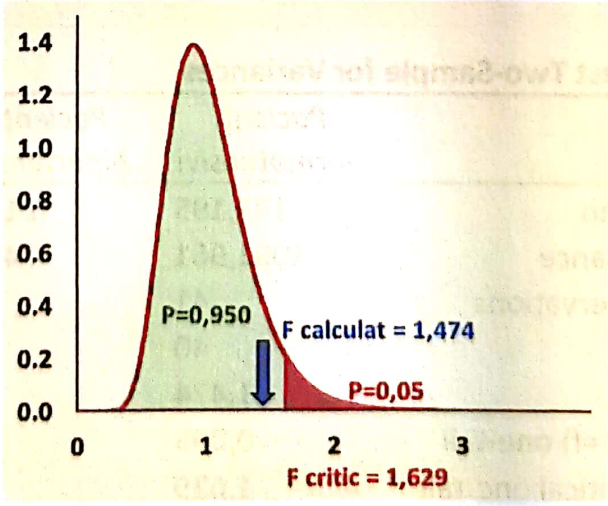


Figura 9.17. Distribuția Fisher cu 40 și 51 grade de libertate

Scanned with OKEN Scanner

Pentru un nivel de semnificație $\alpha = 0{,}05$, se citește valoarea critică din tabelele distribuției Fisher: $F_{critic} = F_{0,05;40;51} = 1{,}629$. Alternativ, putem utiliza formula Excel: $= F.INV.RT(0{,}05; 40; 51)$.

Deoarece $F_{calculat} = 1{,}474 < F_{critic} = 1{,}629$, nu putem respinge ipoteza nulă și prin urmare concluzionăm că nu avem suficiente dovezi pentru a putea considera că varianța în cadrul populației de normotensivi este mai ridicată comparativ cu varianța populației de persoane cu hipertensiune primară, pentru un nivel de semnificație de $\alpha = 5\%$.

În practică, putem utiliza programul Excel în vederea aplicării testului F de testare a raportului variațelor din două populații, astfel: din meniul *„Date"* deschidem modulul de *„Analiză date"* și selectăm opțiunea *„F-Test Two-Sample for Variances"*. În prealabil, informațiile privind variantele variabilelor studiate trebuie prezentate sub forma a două coloane paralele, câte una pentru fiecare variabilă în parte. Prin urmare, este necesar să rearanjăm datele prezentate inițial în tabelul 9.5, în următoarea formă:

Tabelul 9.8. Nivelurile colesterolului (mg/dl) din cele 2 eșantioane, prezentate inițial în tabelul 9.4

Hipertensiune primară	Normotensivi
207	200
172	135
191	161
221	203
...	...

Rulând analiza în programul Excel, vom obține aceleași rezultate:

F-Test Two-Sample for Variances

	Pacienți normotensivi	*Pacienți cu hipertensiune*
Mean	193,195	211,923
Variance	1091,961	740,857
Observations	41	52
df	40	51
F	1,474	
P(F<=f) one-tail	0,095	
F Critical one-tail	1,629	

Scanned with OKEN Scanner

Programul a determinat și probabilitatea de a obține o valoare cel puțin la fel de „extremă" cum este cea calculată la nivelul eșantioanelor noastre, când ipoteza nulă este adevărată, respectiv 0,095 (sau 9,5%). Întrucât probabilitatea $p = 0{,}095$ este mai ridicată în comparație cu nivelul de semnificație $\alpha = 0{,}05$, fixat la începutul studiului, concluzionăm faptul că nu avem suficiente dovezi pentru a putea respinge ipoteza nulă. Prin urmare nu vom respinge ipoteza nulă H_0, $\sigma_1^2 \leq \sigma_2^2$, și vom concluziona că diferența dintre variațele nivelului colesterolului populațiilor de normotensivi și hipertensivi, observată la nivelul celor două eșantioane de lucru ($s_1^2 = 1091{,}961$ și $s_2^2 = 740{,}857$) nu este suficient de mare pentru a fi semnificativă statistic la nivel populațional. Putem spune că probabilitatea de a respinge ipoteza nulă H_0, atunci când ea este adevărată, este prea mare ($p = 0{,}095$) în comparație cu nivelul de semnificație $\alpha = 0{,}05$ utilizat în realizarea studiului nostru.

Scanned with OKEN Scanner

Capitolul. 10

10. ANALIZA VARIANȚEI - ANOVA

SUBIECTE:

10.1. Definiția și principalele tipuri ale analizei variației
10.2. Prezentarea modelului ANOVA unifactorială
10.3. Condițiile preliminare necesare în vederea aplicării testului ANOVA unifactorială
10.4. Componentele principale ale variabilității populației generale - tabelul de calcul ANOVA
10.5. Testele de comparații multiple – analiza post-hoc

La finalizarea acestui capitol, studentul va fi capabil să:
O10-1: să înțeleagă scopul analizei variației precum și situațiile când se utilizează în practică testul ANOVA
O10-2: să evalueze condițiile preliminare de aplicare ale testului ANOVA unifactorial
O10-3: să parcurgă corect etapele testului ANOVA unifactorial pornind de la ipotezele de lucru
O10-4: să interpreteze corespunzător rezultatele testului ANOVA unifactorial
O10-5: să utilizeze când este necesar testele de comparații multiple (analiza post-hoc), ulterior testului ANOVA unifactorial

10.1. DEFINIȚIA ȘI PRINCIPALELE TIPURI ALE ANALIZEI VARIANȚEI

Analiza variației (ANOVA) este o colecție de modele statistice însoțită de procedurile de estimare asociate acestora, fiind utilizată pentru analiza existenței diferențelor între mediile mai multor grupuri independente (nerelaționate) din cadrul aceleiași populații.

ANOVA a fost dezvoltată în anul 1918 de către matematicianul, statisticianul, biologul și geneticianul Sir Ronald Aylmer Fisher „un geniu care aproape de unul singur a creat bazele științei statistice moderne" (Hald Anders, A History of Mathematical Statistics from 1750 to 1930) și se bazează pe *legea variației totale*, în care varianța observată într-o anumită variabilă este împărțită în componente atribuibile diferitelor surse de variație.

În acest mod, variabilitatea totală a populației studiate poate fi împărțită în 2 componente:

Scanned with OKEN Scanner

- *variația între grupuri* (datorată tratamentului aplicat, factor esențial, care a condus la separarea populației în tipuri calitative principale);
- *variația din interiorul fiecărui grup* (cauzată de factori aleatorii, fiind expresia diferențelor individuale dintre unitățile componente).

În forma sa cea mai simplă, ANOVA oferă un test statistic care evaluează situația în care două sau mai multe medii ale grupurilor în care a fost împărțită o populație sunt egale, reprezentând prin urmare generalizarea testului Student „t" în cazul comparării unui număr mai mare de două medii.

Analiza variației poate fi privită ca reprezentând un test statistic de evaluare a potențialelor diferențe dintre grupurile în care a fost împărțită o anumită variabilă dependentă continuă, în funcție de diferitele categorii pe care le pot înregistra una sau mai multe variabile statistice independente calitative (categoriale). Nivelul de măsurare al variabilelor și ipotezele testului joacă un rol important în analiza variației. În cadrul testului ANOVA, variabila dependentă trebuie să fie continuă (de tip interval sau raport), iar variabila / variabilele independente trebuie să fie categoriale (nominale sau ordinale).

În general, în funcție de designul studiului, ANOVA poate fi utilizată în *trei variante principale*:

a) *ANOVA unifactorială*: avem o singură variabilă dependentă continuă care poate fi împărțită în mai multe categorii în funcție de o singură variabilă independentă categorială sau factor; de exemplu, analiza variației poate examina influența a 3 medicamente asupra unei boli, examinând potențialele diferențe existente între timpii medii de vindecare (variabila dependentă continuă) pentru fiecare tip de medicament (variabilă categorială nominală cu 3 variante: 1, 2 și 3).

b) *ANOVA bifactorială*: avem o singură variabilă dependentă continuă care poate fi împărțită în mai multe categorii în funcție de două variabile independente categoriale; de exemplu, analiza variației poate examina potențialele diferențe existente între timpii medii de vindecare (variabila dependentă continuă) în funcție de:
   - *fiecare tip de medicament* (variabila categorială A sau factorul A, cu 3 variante: 1, 2 și 3) și
   - *sexul persoanelor* (variabila categorială B sau factorul B, cu 2 variante: masculin și feminin).

   Deosebim 2 tipuri de efecte asupra variabilei dependente:
   - *două efecte principale*: evidențiază acțiunea fiecărui factor, considerată separat (A și B)
   - *un efect de interacțiune*: descrie acțiunea combinată a celor 2 factori ($A * B$)

Vorbim de un efect de interacțiune atunci când efectul (influența) unui factor depinde de nivelul celui de al doilea factor. Astfel, interacțiunile indică faptul că diferențele nu sunt uniforme între toate categoriile variabilei

Scanned with OKEN Scanner

independente. De exemplu, medicamentul 1 poate avea timpi de vindecare mai mari în comparație cu medicamentele 2 și 3, dar această diferență ar putea fi mai mare (sau mai mică) în cazul persoanelor de sex feminin în comparație cu cele de sex masculin. Scopul principal al ANOVA cu 2 factori este stabilirea existenței interacțiunii dintre cei 2 factori.

c) *ANOVA multifactorială*: avem o singură variabilă dependentă continuă care poate fi împărțită în mai multe categorii în funcție de trei sau mai multe variabile independente categoriale; de exemplu, analiza varianței poate examina potențialele diferențe existente între timpii medii de vindecare (variabila dependentă continuă) în funcție de:

- *fiecare tip de medicament* (variabila categorială A cu 3 variante: 1, 2 și 3)
- *sexul persoanelor* (variabila categorială B cu 2 variante: masculin și feminin)
- *grupele de vârstă ale persoanelor* (variabila categorială C cu 3 variante: tineri, adulți, vârstnici).

Și în acest model sunt prezente efectele de interacțiune. În acest caz vom avea un număr de 7 efecte asupra variabilei dependente, după cum urmează:

- trei efecte principale (A, B și C)
- trei efecte de interacțiune de ordinul 2 ($A * B$, $A * C$ și $B * C$)
- un efect de interacțiune de ordinul 3 ($A * B * C$).

Analiza varianței va începe întotdeauna cu evaluarea efectului de interacțiune de ordin superior (în cazul de față 3), iar în situația în care acesta nu este semnificativ se cercetează efectele de interacțiune de ordin inferior (2).

10.2. PREZENTAREA MODELULUI ANOVA UNIFACTORIALĂ

În cadrul capitolului anterior au fost prezentate noțiunile principale referitoare la testarea ipotezei nule H_0 a absenței diferențelor dintre mediile a două populații (testele „z" și „t", în funcție de caz). De multe ori în practică însă suntem interesați să testăm ipoteza nulă a absenței diferențelor existente între mediile mai multor populații. La prima vedere, am putea fi tentați să realizăm o serie de comparații între toate perechile de medii ale eșantioanelor de lucru, în mod independent, prin utilizarea testului Student.

De ce nu este fezabilă compararea grupurilor utilizând mai multe teste Student ?

La fiecare aplicare a unui test Student există o șansă de 5% de a ne confrunta cu o eroare de tip I ($\alpha = 0{,}05$). De exemplu în situația realizării unei comparații între mediile a 4 sub-populații (A, B, C și D) avem un număr de:

Scanned with OKEN Scanner

$$C_4^2 = \frac{4!}{2! \cdot (4-2)!} = 6$$

comparații între fiecare 2 grupuri extrase din acestea ($A-B, A-C, A-D, B-C, B-D$ și $C-D$), adică 6 teste Student „t", fiecare având o eroare de tip I de $\alpha = 0{,}05$.

Probabilitatea de a obține un rezultat corect în cadrul fiecărui test în parte este: $1-\alpha = 0{,}95$. Având în vedere faptul că fiecare dintre aceste 6 teste Student se realizează în mod independent unul de celălalt, probabilitatea de a obține un rezultat corect pentru întregul set de 6 teste, va fi produsul acestor probabilități (regula specială a înmulțirii probabilităților în cazul evenimentelor independente – prezentată anterior în capitolul 4.3), respectiv:

$$0{,}95 \cdot 0{,}95 \cdot 0{,}95 \cdot 0{,}95 \cdot 0{,}95 \cdot 0{,}95 = 0{,}735.$$

În consecință, eroarea de tip I aferentă aplicării unui set de 6 teste Student succesive este $\alpha = 1 - 0{,}735 = 0{,}265$ sau 26,5%, valoare care nu este acceptabilă în cadrul niciunei analize statistice practice (riscul de a respinge o ipoteză nulă H_0, care este de fapt adevărată, a crescut foarte mult, depășind nivelul maxim acceptat în cadrul majorității studiilor științifice de 5%).

Rezultă imediat faptul că *este necesară utilizarea unui singur test statistic* (Fisher „F") care să compare simultan mediile sub-populațiilor de interes și să păstreze în acest fel eroarea de tip I (α) sub nivelul prag de 5%. Varianta cea mai simplă de analiză a variației este cea unifactorială, în care cercetătorul investighează doar o singură sursă de variație sau factor, putând fi astfel privită ca o extensie a testului Student „t" în cazul studierii unui număr de trei sau mai multe eșantioane independente.

Sintetizând, putem afirma faptul că ANOVA este o procedură de analiză și evaluare a existenței potențialelor diferențe între mediile mai multor sub-populații independente (nerelaționate), care aparțin aceleiași populații generale, de interes pentru o anumită cercetare. În cadrul acestui model avem o singură variabilă dependentă continuă, împărțită în mai multe clase (grupuri), în funcție de nivelurile unei singure variabile independente de tip categorial (factor).

Ipotezele modelului ANOVA unifactorială:

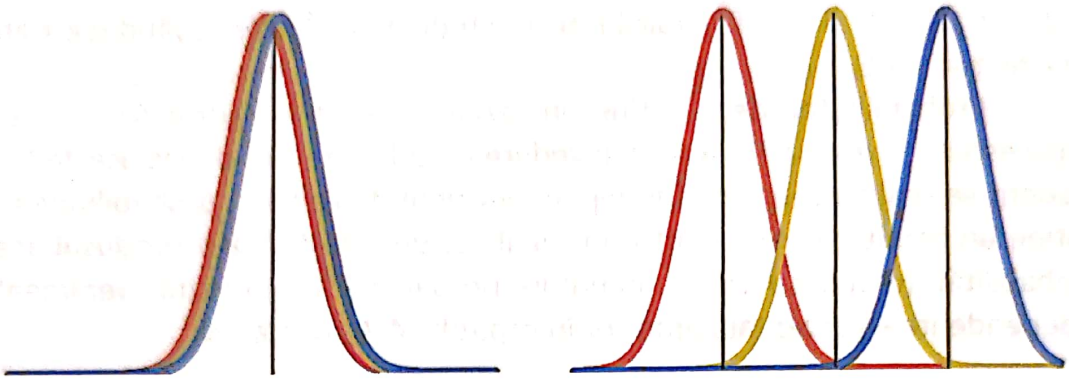
- *Ipoteza nulă* (H_0) pentru testul ANOVA: nu există nicio diferență semnificativă între mediile sub-populațiilor comparate. În situația când avem k grupuri, iar μ_i este media sub-populației grupului j, $j = \overline{1,k}$:

$$\mu_1 = \mu_2 = \mu_3 = \cdots = \mu_j = \cdots = \mu_k$$

- *Ipoteza alternativă* (H_1) presupune că există cel puțin o diferență semnificativă între mediile sub-populațiilor din care provin grupurile

Scanned with OKEN Scanner

investigate, fără însă să localizeze în mod explicit diferența. Pentru a determina mediile căror sub-populații diferă semnificativ între ele, ulterior efectuării analizei variației trebuie aplicate testele post-hoc.



H_0 adevărată ($\mu_1 = \mu_2 = \mu_3$) H_0 falsă ($\mu_1 \neq \mu_2 \neq \mu_3$)

Figura 10.1. Ipotezele testului ANOVA unifactorială în cazul a 3 sub-populații normal distribuite care au variațele egale ($\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2$)

În practică, ANOVA unifactorială poate fi utilizată în cadrul a două tipuri principale de studii, în funcție de modul în care sunt alocați participanții în grupuri și de felul în care sunt manipulate variabilele independente (factorii):

1) *loturi distincte realizate aleatoriu, pe baza unui factor experimental*: se recrutează un eșantion de persoane, care este apoi împărțit în mod aleatoriu în trei sau mai multe grupuri, cu condiția ca fiecare participant să fie alocat unui singur grup (tipul de tratament, tipul de dietă, tipul de exerciții, etc.); ulterior, persoanele din fiecare grup sunt supuse unor sarcini diferite și este măsurat apoi rezultatul înregistrat de fiecare individ pentru variabila dependentă de interes în cadrul studiului (tensiunea arterială, nivelul colesterolului, timpul de recuperare, scorul pe scala de depresie, rata infecțiilor după vaccinare, etc.); în cadrul acestui tip de studiu, variabila independentă este manipulată activ de cercetător pe parcursul derulării experimentului; alocarea aleatorie ajută în acest caz la controlul influenței variabilelor externe, minimizând în consecință efectul altor factori care ar putea denatura rezultatele studiului; din acest considerent, designul 1 poate fi considerat util în vederea testării ipotezelor cu privire la cauzalitatea dintre variabile.

În continuare, prezentăm o posibilă modalitate de alocare în mod aleatoriu a pacienților din cadrul unui eșantion de 15 persoane în trei grupuri de tratament, prin utilizarea programului Excel. Fiecărui pacient extras din carul populației generale (folosind modulul Excel „Data Analysis"), i s-a atribuit în mod aleatoriu un număr cuprins între 0 și 1, prin utilizarea funcției Excel: $RAND()$.

Scanned with OKEN Scanner

Realizând o simplă ordonare, de la valoarea minimă către valoarea maximă, a celor 15 numere aleatorii alocate pacienților din eșantion, au rezultat cele trei grupuri de pacienți, fiecare având același volum (5 persoane).

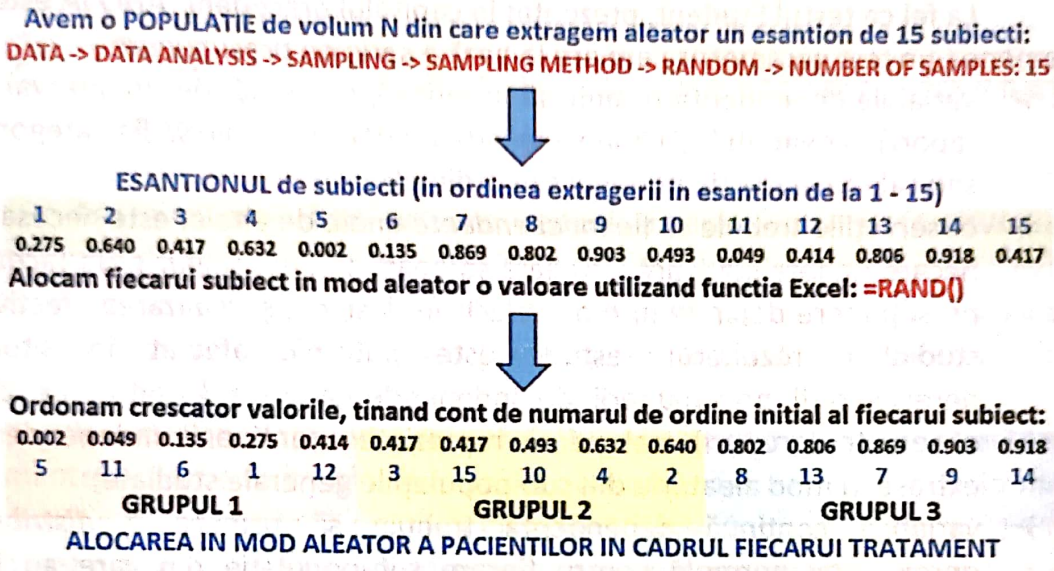

Avem o POPULATIE de volum N din care extragem aleator un esantion de 15 subiecti:
DATA -> DATA ANALYSIS -> SAMPLING -> SAMPLING METHOD -> RANDOM -> NUMBER OF SAMPLES: 15
ESANTIONUL de subiecti (in ordinea extragerii in esantion de la 1 - 15)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
0.275 0.640 0.417 0.632 0.002 0.135 0.869 0.802 0.903 0.493 0.049 0.414 0.806 0.918 0.417
Alocam fiecarui subiect in mod aleator o valoare utilizand functia Excel: =RAND()
Ordonam crescator valorile, tinand cont de numarul de ordine initial al fiecarui subiect:
0.002 0.049 0.135 0.275 0.414 0.417 0.417 0.493 0.632 0.640 0.802 0.806 0.869 0.903 0.918
5 11 6 1 12 3 15 10 4 2 8 13 7 9 14
GRUPUL 1
GRUPUL 2
GRUPUL 3
ALOCAREA IN MOD ALEATOR A PACIENTILOR IN CADRUL FIECARUI TRATAMENT

Figura 10.2. Extragerea unui eșantion de 15 pacienți din cadrul unei populații și împărțirea în mod aleatoriu a acestora în 3 grupuri de tratament prin utilizarea programului Excel

2) *loturi în care participanții sunt deja împărțiți pe baza unui factor preexistent*: se extrage în mod aleatoriu un eșantion de persoane, care este ulterior împărțit în grupuri în funcție de valorile unei variabile categoriale independente care reprezintă o caracteristică naturală, obiectivă a subiecților (grupa de vârstă, grupa de sânge, statutul de fumător, nivelul activității fizice, etc.), fiecare individ putând aparține în mod evident unui singur grup; persoanele din cadrul grupurilor sunt măsurate în funcție de o variabilă dependentă (greutatea, tensiunea arterială, nivelul colesterolului, timpul de recuperare, nivelul de anticorpi, etc.), după ce au îndeplinit aceleași sarcini sau au fost supuse acelorași condiții; în acest tip de studiu, variabila independentă este observată de cercetător și reflectă o caracteristică naturală a participanților, neexistând prin urmare o manipulare experimentală precum în cazul anterior; în acest caz, nu există un control al variabilelor externe, deoarece nu s-a realizat o alocare aleatorie a subiecților în cadrul grupurilor (de exemplu, grupa de vârstă poate influența nivelul colesterolului, dar cercetătorul nu poate interveni asupra acestui factor).

Scanned with OKEN Scanner

10.3. CONDIȚIILE PRELIMINARE NECESARE ÎN VEDEREA APLICĂRII TESTULUI ANOVA UNIFACTORIAL

La fel ca testul Student, prezentat în capitolul precedent, ANOVA este de asemenea un test parametric care are la bază o serie de *presupuneri*:

- variabila dependentă trebuie să numerică, continuă (de tip interval sau raport), iar variabila / variabilele independente trebuie să fie categoriale sau calitative (de tip nominal sau ordinal);
- observațiile trebuie să fie *independente* unele de altele: este necesar ca fiecare unitate statistică studiată să aparțină unui singur grup (această presupunere depinde în mod direct de designul și realizarea efectivă a studiului); rezultatul testului este puternic afectat în situația nerespectării presupunerii de independență; cele k seturi de date observate (grupuri) trebuie să reprezinte eșantioane independente, extrase în mod aleatoriu din sub-populațiile generale studiate;
- variabila continuă dependentă trebuie să urmeze o *distribuție aproximativ normală* pentru fiecare sub-populație din care au fost extrase grupurile sau eșantioanele supuse comparației: în acest sens va fi realizată inițial o evaluare grafică (histograma, box-plot și QQ plot), urmată de calcularea indicatorilor statistici descriptivi (medie, mediană, mod, varianță, asimetrie și boltire) iar, în final for fi utilizate și testele de evaluare a normalității (Shapiro-Wilk și Kolmogorov-Smirnov); rezultatele sunt rezonabile și în situația nerespectării normalității distribuțiilor populațiilor din care provine fiecare grup, cu condiția utilizării unor eșantioane de volum mare ($n_i \geq 30$);
- variațele sub-populațiilor din care provine fiecare grup trebuie să fie aproximativ egale $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \cdots = \sigma_j^2 = \cdots = \sigma_k^2 = \sigma^2$, unde $j = \overline{1,k}$ (*omogenitatea variației*): în acest sens, se utilizează testele Levene, Brown-Forsythe sau Bartlett; respectarea acestei presupuneri este esențială, mai ales în situația utilizării testului ANOVA asupra unor grupuri (eșantioane) de volum foarte diferit; în cazul nerespectării acestei presupuneri, rezultatele ANOVA sunt considerate relativ robuste dacă sunt utilizate în analiză eșantioane de același volum;
- variabila dependentă nu prezintă valori extreme semnificative (outlieri) în cadrul niciunui grup analizat (valorile extreme prezente la nivel de eșantion pot distorsiona normalitatea distribuției datelor studiate, denaturând diferențele reale dintre mediile grupurilor comparate sau crescând în mod artificial varianța unuia sau mai multor grupuri; toate aceste aspecte pot conduce în final la obținerea unor rezultate incorecte în cazul testului ANOVA, afectând în final validitatea analizei;

Scanned with OKEN Scanner

- *erorile* (diferențele dintre fiecare valoare individuală și media grupei căreia îi aparține) sunt *independente* și urmează o *distribuție normală*, având media egală cu valoarea zero; normalitatea se evaluează în mod formal la nivelul erorilor (valorilor reziduale), pentru întregul set de date.

10.4. COMPONENTELE PRINCIPALE ALE VARIABILITĂȚII POPULAȚIEI GENERALE – TABELUL DE CALCUL ANOVA

Așa cum am văzut anterior, scopul analizei variației (ANOVA) este reprezentat de compararea mediilor mai multor grupuri pentru a putea evalua dacă diferențele observate pot fi atribuite unui efect real, cauzat de aplicarea unui tratament, sau pot fi atribuite doar unei simple variații întâmplătoare (sunt datorate erorii).

În construirea modelului ANOVA unifactorială, pentru a evidenția elementele componente ale unei valori observate x_{ij} vom porni de la *media generală* μ (valoarea tipică obținută în condiții ideale, în cazul în care toți indivizii ar fi fost tratați la fel, neexistând astfel tratamente sau grupuri diferite), la care vom adăuga *efectul tratamentului* efectuat α_j (cât de mult diferă sub-populația j, în medie, față de media generală) și apoi *termenul de eroare* ε_{ij} (diferența dintre valoarea observată x_{ij} și valoarea așteptată pentru sub-populația respectivă, sau media sub-populației j). Astfel, avem:

$$x_{ij} = \mu + \alpha_j + \varepsilon_{ij} = \mu + (\mu_j - \mu) + (x_{ij} - \mu_j), \tag{10.1}$$

unde:

μ = media globală sau generală (media întregii populații)

$\alpha_j = \mu_j - \mu$ (efectul tratamentului efectuat)

$\varepsilon_{ij} = x_{ij} - \mu_j$ (eroarea pentru observația i din sub-populația j)

Formula 10.1 se obține pornind de la relația identitară: $x_{ij} = x_{ij}$, la care adăugăm și scădem media populației generale (μ) și media sub-populației j (μ_j), din care face parte unitatea noastră statistică având varianta x_{ij} a variabilei dependente studiate X.

Pentru a contura o imagine mai clară asupra ANOVA unifactorială, datele colectate inițial la nivelul eșantionului de lucru se pot sistematiza într-un tabel cu dublă intrare, în care pe coloane se înscriu cele k tratamente (grupuri), iar pe rânduri variantele variabilei dependente din fiecare grup (tabelul 10.1). Considerăm, pentru simplificare, faptul că, fiecare grup din cel k grupuri în care a fost împărțit eșantionul de lucru are același volum n. Prin urmare, *volumul întregului eșantion*, divizat în k grupuri, fiecare de volum n, va fi: $N = n \cdot k$

Scanned with OKEN Scanner

Tabelul 10.1. Tabelul cu dublă intrare utilizat în ANOVA unifactorială

Variantele variabilei X	Tratamentul (grupul)						Total
	1	2	...	j	...	k	
1	x_{11}	x_{12}	...	x_{1j}	...	x_{1k}	$\sum_{j=1}^{k} x_{1j}$
2	x_{21}	x_{22}	...	x_{2j}	...	x_{2k}	$\sum_{j=1}^{k} x_{2j}$
...	...	...	...	...	...	...	
i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{ik}	$\sum_{j=1}^{k} x_{ij}$
...	...	...	...	...	...	...	
n	x_{n1}	x_{n2}	...	x_{nj}	...	x_{nk}	$\sum_{j=1}^{k} x_{nj}$
Total	$\sum_{i=1}^{n} x_{i1}$	$\sum_{i=1}^{n} x_{i2}$	...	$\sum_{i=1}^{n} x_{ij}$	...	$\sum_{i=1}^{n} x_{ik}$	$\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}$
Volum	n	n	...	n	...	n	$n \cdot k$
Media	$\bar{x}_1$	$\bar{x}_2$	...	$\bar{x}_j$	...	$\bar{x}_k$	$\bar{x}$

În majoritatea cazurilor, cele k tratamente (grupuri) reflectă influența unui anumit factor (variabilă independentă categorială) care acționează asupra variabilei dependente studiate. Astfel se separă influența factorilor esențiali, de influențele celorlalți factori, considerați întâmplători.

Media fiecărui grup j (tratament) se determină astfel:

$$\bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n}$$

în care:

x_{ij} reprezintă varianta i a variabilei X aflată în grupul j

n este volumul sau numărul de unități statistice din grupul j (în cazul nostru, toate cele j grupuri au același volum n)

Media generală $\bar{x}$ a eșantionului se poate determina în două modalități:

Scanned with OKEN Scanner

- suma tuturor celor $N = n \cdot k$ variante ale variabilei dependente X, raportată la volumul total al eșantionului N:

$$\bar{x} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}}{N} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}}{n \cdot k} \tag{10.2}$$

- media mediilor celor k grupuri în care a fost împărțit eșantionul:

$$\bar{x} = \frac{\sum_{j=1}^{k} \bar{x}_j}{n \cdot k} = \frac{\sum_{j=1}^{k} \frac{\sum_{i=1}^{n} x_{ij}}{n}}{n \cdot k} = \frac{\sum_{j=1}^{k} \left(\frac{1}{n} + \frac{1}{n} + \ldots + \frac{1}{n}\right) \sum_{i=1}^{n} x_{ij}}{n \cdot k} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k} x_{ij}}{n \cdot k} \tag{10.3}$$

ANOVA unifactorială presupune descompunerea variației totale a variabilei studiate X în două componente principale (în interiorul fiecărui grup și între grupuri), estimarea varianțelor corespunzătoare pentru fiecare componentă și, în final, compararea acestora sub forma unui raport.

Variația (abaterea) fiecărei variante a variabilei x_{ij} față de media generală $\bar{x}$ a eșantionului se poate descompune în 2 componente, prin scăderea și adunarea mediei grupului căreia îi aparține aceasta ($\bar{x}_j$):

$$x_{ij} - \bar{x} = (x_{ij} - \bar{x}_j) + (\bar{x}_j - \bar{x}) \tag{10.4}$$

în care:

$x_{ij} - \bar{x}_j$ reprezintă abaterea observației i din grupul j față de media grupului j, evidențiind nivelul variabilității interne (intra-grup);

$\bar{x}_j - \bar{x}$ reprezintă abaterea mediei grupului j față de media generală, fiind o expresie a variabilității existente între grupuri (inter-grup).

Ridicând la pătrat relația (10.4) obținem:

$$(x_{ij} - \bar{x})^2 = (x_{ij} - \bar{x}_j)^2 + 2 \cdot (x_{ij} - \bar{x}_j) \cdot (\bar{x}_j - \bar{x}) + (\bar{x}_j - \bar{x})^2 \tag{10.5}$$

Însumăm relația 10.5 pentru toate observațiile din eșantion și efectuăm calculele:

$$\sum_{i=1}^{n}\sum_{j=1}^{k}(x_{ij} - \bar{x})^2 = \sum_{i=1}^{n}\sum_{j=1}^{k}(x_{ij} - \bar{x}_j)^2 + 2\sum_{i=1}^{n}\sum_{j=1}^{k}(x_{ij} - \bar{x}_j) \cdot (\bar{x}_j - \bar{x}) + \sum_{i=1}^{n}\sum_{j=1}^{k}(\bar{x}_j - \bar{x})^2 \tag{10.6}$$

Deoarece $\bar{x}_j - \bar{x}$ este o constantă în cadrul fiecărui grup j, nedepinzând în niciun fel de valorile lui i, vom avea:

Scanned with OKEN Scanner

$$2\sum_{i=1}^{n}\sum_{j=1}^{k}(x_{ij}-\bar{x}_j)\cdot(\bar{x}_j-\bar{x})=2\sum_{j=1}^{k}(\bar{x}_j-\bar{x})\sum_{i=1}^{n}(x_{ij}-\bar{x}_j) \tag{10.7}$$

Conform proprietăților mediei aritmetice, prezentate anterior în capitolul al treilea al cărții, cunoaștem faptul că suma abaterilor fiecărui nivel individual de la nivelul mediu al seriei este zero. Prin urmare, la nivelul fiecărei grupe j din eșantion vom avea:

$$\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)=0 \tag{10.8}$$

Înlocuind relația 10.8 în 10.7, observăm faptul că, termenul din mijloc al relației 10.6 se anulează:

$$2\sum_{j=1}^{k}(\bar{x}_j-\bar{x})\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)=2\cdot 0\cdot\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)=0$$

În consecință, relația 10.6 poate fi scrisă astfel:

$$\sum_{i=1}^{n}\sum_{j=1}^{k}(x_{ij}-\bar{x})^2=\sum_{i=1}^{n}\sum_{j=1}^{k}(x_{ij}-\bar{x}_j)^2+\sum_{i=1}^{n}\sum_{j=1}^{k}(\bar{x}_j-\bar{x})^2 \tag{10.9}$$

Deoarece ultima sumă din relația (10.9) depinde doar de numărul de grupe (j), relația anterioară devine:

$$\sum_{i=1}^{n}\sum_{j=1}^{k}(x_{ij}-\bar{x})^2=\sum_{i=1}^{n}\sum_{j=1}^{k}(x_{ij}-\bar{x}_j)^2+n\cdot\sum_{j=1}^{k}(\bar{x}_j-\bar{x})^2 \tag{10.10}$$

În acest mod, suma pătratelor abaterilor totale SST este împărțită în 2 componente suma pătratelor abaterilor din interiorul fiecărui grup (reziduale) SSW și suma pătratelor abaterilor între grupuri SSB (datorată factorilor esențiali, care au condus la împărțirea colectivității în tipuri mari calitative):

$$SST=SSW+SSB \tag{10.11}$$

Considerând faptul că cele k grupuri din componența eșantionului de lucru ar fi fost extrase din aceeași populație generală, varianța acestei populații poate fi estimată în trei feluri:

Scanned with OKEN Scanner

- *varianța reziduală*, calculată prin raportarea sumei pătratelor abaterilor din interiorul fiecărui grup (SSW), la numărul gradelor de libertate asociate acestora (df_{within});
- *varianța factorială ajustată*, calculată prin împărțirea sumei pătratelor abaterilor între grupuri (SSB), la numărul de grade de libertate aferente ($df_{between}$);
- *varianța totală*, determinată prin împărțirea sumei pătratelor abaterilor totale (SST), la numărul total de grade de libertate (df_{total}).

În continuare, vom determina numărul gradelor de libertate, separat pentru fiecare dintre cele 3 sume de pătrate evidențiate anterior, în relațiile 10.10 și 10.11. Numărul total de variante ale variabilei dependente (N) este egal cu numărul de grupe (k) înmulțit cu numărul de variante repartizate în fiecare grupă (n):

$$N = k \cdot n$$

Prin urmare, pentru suma pătratelor abaterilor totale a celor N variante x_{ij} (SST), vom avea:

$$df_{total} = N - 1 = k \cdot n - 1 \tag{10.12}$$

Similar cu calculul varianței unui eșantion, pentru a estima varianța populației generale din care acesta provine, am pierdut un grad de libertate din totalul de N, pentru a estima media necunoscută μ a populației pe baza mediei calculate la nivel de eșantion $\bar{x}$.

Gradele de libertate (df_{within}) asociate cu suma pătratelor abaterilor în interiorul fiecărei grupe (SSW). Aceasta s-a determinat prin însumarea sumelor pătratelor abaterilor din interiorul fiecărui grup ($SS_1 + SS_2 + \cdots + SS_j + \cdots + SS_k$). Fiecare sumă a unui grup SS_j are asociat un număr de $n - 1$ grade de libertate. Rezultă că:

$$df_{within} = (n-1) + (n-1) + \cdots + (n-1) = k \cdot (n-1) = \\ = k \cdot n - k = N - k$$

Rezultă că:

$$df_{within} = N - k \tag{10.13}$$

Numărul gradelor de libertate ($df_{between}$) asociat cu suma pătratelor abaterilor între grupe (SSB), calculat ca suma pătratelor diferențelor dintre mediile celor k grupuri și media generală, va fi:

$$df_{between} = k - 1 \tag{10.14}$$

Scanned with OKEN Scanner

Suma gradelor de libertate ale celor două componente este egală cu numărul total de grade de libertate:

$$df_{total} = df_{within} + df_{between} \tag{10.15}$$

În situația în care presupunerile modelului sunt validate, iar mediile populațiilor din care au fost extrase grupurile sunt egale (ipoteza nulă H_0 este adevărată), ambele sume de pătrate (SSW și SSB) raportate la gradele de libertate corespunzătoare, reprezintă estimatori independenți și nedeplasați ai varianței populației generale.

Prin urmare, vom avea:

- *varianța reziduală*:

$$MSW = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k}\left(x_{ij} - \bar{x}_j\right)^2}{N-k} \tag{10.16}$$

Varianța intra-grup sau reziduală (MSW) ne oferă un estimator nedeplasat al varianței reale a populației generale, doar în situația în care varianțele celor j sub-populații sunt egale. Nu este necesar ca ipoteza nulă H_0 a egalității mediilor celor j sub-populații să fie adevărată, atât timp cât varianțele acestora sunt egale.

- *varianța factorială*:

$$MSB = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k}\left(x_{ij} - \bar{x}_j\right)^2}{k-1} \tag{10.17}$$

Varianța inter-grupuri sau factorială (MSB) reprezintă un o estimare validă a varianței populației generale în situația în care atât varianțele celor j sub-populații sunt egale cât și ipoteza nulă H_0 a egalității mediilor celor j sub-populații este adevărată.

Prin intermediul testului F vom evalua raportul acestor două tipuri de variație (între grupuri și intra-grup). O valoare ridicată a acestui raport sugerează că variația între grupuri este semnificativ mai mare decât variația intra-grup, ceea ce indică faptul că diferențele între mediile grupurilor nu sunt întâmplătoare și că tratamentul aplicat a condus la separarea populației în tipuri calitative distincte.

Raportul mediilor acestor două surse de variație urmează o *distribuție Fisher*, cu $k-1$ grade de libertate la numărător, respectiv $N-k$ grade de libertate la numitor:

$$F_{calculat} = \frac{MSB}{MSW} \tag{10.18}$$

Scanned with OKEN Scanner

În situația în care cele două estimări a varianței populației generale (MSW și MSB) sunt aproximativ egale, raportul lor va fi foarte apropiat de valoarea 1, susținând ipoteza nulă a egalității mediilor sub-populațiilor din care provin grupurile. Pe de altă parte, în cazul în care varianța inter-grupuri este mult mai ridicată comparativ cu varianța intra-grup, raportul lor va fi și el supraunitar. Prin urmare, o valoare suficient de mare în raport cu 1 a raportului celor două varianțe va genera o umbră de neîncredere asupra ipotezei nule H_0 a egalității mediilor celor j sub-populații. Întrebarea care se pune este în acest context este: *ce valoare supraunitară trebuie să înregistreze acest raport al varianțelor pentru a putea fi dispuși să concluzionăm că diferența observată la nivel de eșantion, între cei doi estimatori ai varianței populației generale, nu este obținută din pură întâmplare (nu este datorată doar șansei)?*.

În cadrul analizelor practice, valoarea calculată $F_{calculat}$ a raportului varianțelor obținută la nivel de eșantion se compară cu valoarea din tabelul distribuției Fisher, corespunzătoare gradelor de libertate și unui nivel de semnificație standard ($\alpha = 0{,}05$): $F_{critic} = F_{\alpha,k-1,N-k}$ (valoarea critică). În programul Excel putem utiliza funcția: $= F.INV.RT(0{,}05;\ df1;\ df2)$, care returnează același rezultat.

Dacă $F_{calculat} \geq F_{critic}$ ($p < 0{,}05$) raportul varianțelor se situează în zona de respingere, în extremitatea dreaptă a distribuției (figura 9.14) și, în consecință, vom respinge ipoteza nulă a egalității mediilor populațiilor.

În cazul în care $F_{calculat} < F_{critic}$ ($p > 0{,}05$) nu avem suficiente dovezi pentru a respinge H_0, diferența dintre cele două estimări ale varianței populației putând fi datorată șansei.

Datele specifice ANOVA unifactorială sunt prezentate în majoritatea programelor de analiză statistică într-un tabel care are următoarea structură:

Tabelul 10.2. Tabelul ANOVA unifactorială

Sursa variației	**Suma pătratelor**	**Gradele de libertate**	**Media pătratelor**	**Raportul varianțelor (Fisher)**
- între grupe	SSB	$k-1$	$MSB = \frac{SSB}{k-1}$	$F_{calculat} = \frac{MSB}{MSW}$
- în interiorul grupelor	SSW	$N-k$	$MSW = \frac{SSW}{N-k}$	-
Total	SST	$N-1$	-	-

Respingerea ipotezei asupra celor două varianțe implică în același timp și respingerea ipotezei nule H_0 a egalității mediilor celor j sub-populații. Întrucât

Scanned with OKEN Scanner

varianța inter-grupuri de la numărătorul formulei are la bază împrăștierea mediilor grupurilor din eșantion în jurul mediei generale, această valoare va fi cu atât mai ridicată cu cât diferențele dintre mediile grupurilor vor fi mai mari. Cu alte cuvinte, o valoare suficient de mare a acestui raport de varianțe ne va sugera posibilitatea respingerii ipotezei nule H_0 a egalității mediilor sub-populațiilor, din care provin grupurile studiate.

În concluzie, atunci când respingem ipoteza nulă H_0 concluzionăm că nu toate mediile populațiilor sunt egale, fără a putea preciza însă perechile de medii care sunt diferite. În situația în care nu reușim să respingem ipoteza nulă, considerăm că mediile populațiilor studiate nu sunt semnificativ diferite între ele, eventualele diferențe evidențiate la nivel de eșantion fiind cel mai probabil datorate șansei.

Exemplu: într-un studiu, un cercetător dorește să determine influența a trei medicamente asupra unei afecțiuni. În acest sens, a fost selectat în mod aleatoriu un eșantion de 63 subiecți diferiți din populația generală, care suferă de acea afecțiune, din care au fost repartizați, în mod cu totul întâmplător, câte 21 de subiecți pentru fiecare dintre cele 3 tratamente (tabelul 10.3). În vederea evaluării eficienței medicamentelor s-a măsurat perioada de timp exprimată în săptămâni necesară fiecărui pacient pentru a se însănătoși. Pentru testarea ipotezei nule a egalității mediilor, în situația studierii a 3 sau mai multe populații, se va utiliza testul ANOVA unifactorială.

Tabelul 10.3. Timpul de însănătoșire a pacienților (săptămâni), în două variante de prezentare, în funcție de aplicația utilizată pentru analiză

Med. 1	Med. 2	Med. 3
1	8	1
3	7	3
5	5	4
4	6	2
7	9	6
6	7	3
7	2	5
8	7	1
4	9	7
5	8	5
...	...	...

Varianta 1 (Excel)

Medicament	Timp însănătoșire
1	1
1	3
1	5
...	...
2	8
2	7
2	5
...	...
3	1
3	3
3	4

Varianta 2 (SPSS)

Scanned with OKEN Scanner

Evaluarea presupunerilor modelului:

a) *sub-populațiile din care au fost extrase în mod aleatoriu cele trei grupuri sunt aproximativ normal distribuite*

Realizăm la început o prezentare a principalilor indicatori statistici descriptivi, alături de o reprezentare grafică, pentru a contura o primă imagine asupra informațiilor din eșantionul studiat:

Descriptive statistics	Med. 1	Med. 2	Med. 3
No. of values	21	21	21
Mean	5,190	7,333	4,190
Median	5,00	7,00	4,00
Variance	2,962	2,533	3,762
Std. Deviation	1,721	1,592	1,940
Minimum	2	4	1
Maximum	8	10	8
Skewness	0,065	-0,198	0,158
Kurtosis	-0,964	-0,202	-0,551

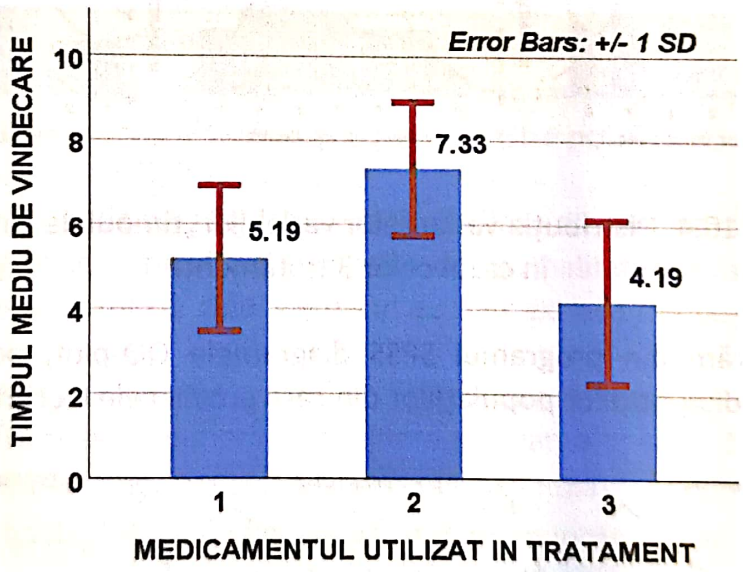


Figura 10.3. Timpii medii de vindecare în cazul celor 3 tratamente

Medicamentul 2 a avut ca efect cea mai lungă perioadă medie de vindecare, de peste 7 săptămâni. Medicamentul 3 a determinat cel mai scurt timp de vindecare (aproximativ 4 săptămâni), în timp ce medicamentul 1 condus la o perioadă medie de timp până la însănătoșire situată între celelalte 2 (aproximativ 5 săptămâni). Observăm că nivelul de variabilitate, determinat pe baza datelor din eșantion este relativ similar în cazul celor trei grupuri. De

Scanned with OKEN Scanner

asemenea, indicatorii tendinței centrale (media și mediana) înregistrează valori apropiate pentru fiecare grup. În cazul distribuțiilor normale, valorile coeficienților de asimetrie și boltire ar trebui să fie apropiate de zero, valorile care nu se încadrează în intervalul $(-1,1)$ sugerând abateri de la normalitate (asimetrie la stânga/dreapta și/sau aplatizare/boltire mai mari decât în cazul distribuției gausiene). În cazul de față, coeficienții de asimetrie sunt: 0,065; -0,198 și 0,158, iar coeficienții de boltire sunt -0,964; -0,202 și -0,551, toate aceste valori sugerând abateri mici de la normalitate.

Realizăm în continuare *histogramele celor trei grupuri*, pentru a observa vizual distribuțiile de frecvență ale variabilei „timpul de vindecare".

În situația în care histogramele au formă de clopot și sunt relativ simetrice ($\bar{x} = Me$), vom putea considera faptul că, variabila statistică analizată urmează distribuții aproximativ normale, în cazul celor trei tipuri de tratament (figura 10.4):

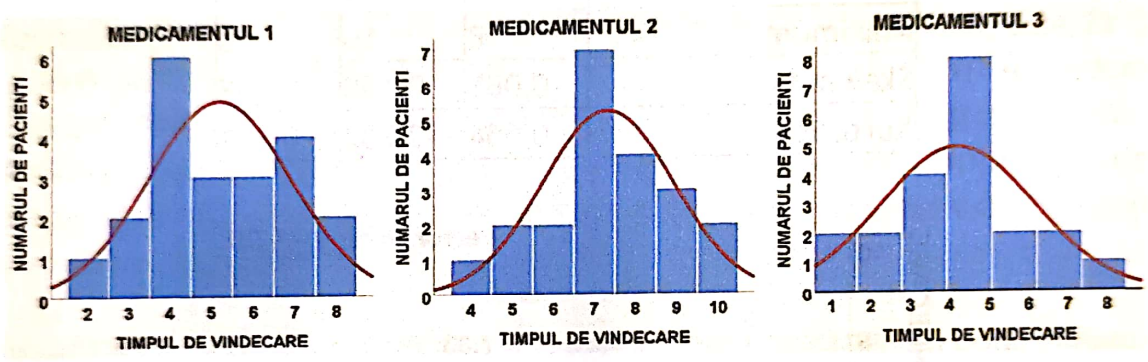


Figura 10.4. Distribuția variantelor variabilei „timpul de vindecare" în cazul celor 3 tratamente

Generăm din programul SPSS diagramele QQ-plot, pentru a evalua normalitatea distribuțiilor populațiilor din care provin cele trei grupuri

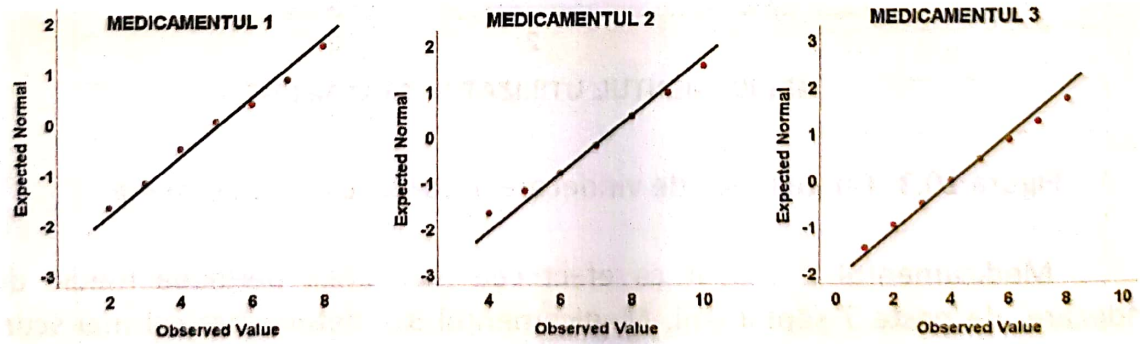


Figura 10.5. Diagramele Q-Q Plot ale variabilei „timpul de vindecare" în cazul celor 3 tratamente

Scanned with OKEN Scanner

Analizând datele prezentate în figura 10.5, observăm faptul că punctele nu deviază foarte mult de la liniile de referință înclinate la 45^o și, în consecință, putem concluziona că grupurile provin din populații care urmează distribuții aproximativ normale.

Vom verifica ulterior prin teste statistice normalitatea populațiilor din care au fost extrași în mod aleatoriu pacienții aparținând celor trei grupuri de tratament:

Tests of Normality

		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	MEDICAMENT	Statistic	df	Sig.	Statistic	df	Sig.
TIMP_VINDECARE	1	.184	21	.062	.941	21	.230
	2	.179	21	.078	.953	21	.384
	3	.111	21	.200*	.966	21	.655

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Valorile testelor de normalitate Kolmogorov-Smirnov și Shapiro-Wilk realizate în SPSS evidențiază faptul că nu avem suficiente dovezi pentru a respinge ipoteza nulă a normalității populațiilor din care provin eșantioanele studiate (valorile probabilităților p aferente testelor de normalitate sunt în toate cazurile $p \geq 0{,}05$). Cu alte cuvinte, putem considera faptul că sub-populațiile din care au fost extrase în mod aleatoriu grupurile noastre pot fi aproximativ normal distribuite.

Având în vedere evaluarea indicatorilor de statistică descriptivă, reprezentările grafice și rezultatele testelor de normalitate, detaliate anterior, putem presupune faptul că, cele 3 grupuri au fost extrase din populații normal distribuite.

b) populațiile au varianțe egale (condiția de homoscedasticitate)

În vederea verificării presupunerii de egalitate a varianțelor celor 3 sub-populații vom utiliza testul Levene. Acesta testează ipoteza nulă a egalității varianțelor celor trei sub-populații H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$, împotriva ipotezei alternative H_1: cel puțin două sub-populații au varianțe diferite. În analiza practică, se va determina media fiecărui grup din eșantion ($\bar{x}_j$), se va calcula deviația absolută a fiecărei observații față de aceasta ($x_{ij} - \bar{x}_j$), iar în final se va aplica o ANOVA asupra acestor deviații absolute. În cazul existenței unor abateri mai importante de la normalitate (existența unei asimetrii puternice sau prezența unor valori extreme), se vor determina deviațiile absolute față de mediana fiecărui grup ($x_{ij} - Me$), testul fiind mai robust în acest caz comparativ cu situația utilizării mediei. Calculele pot fi realizate etapizat în Microsoft Excel, sau se poate folosi programul SPSS, care generează automat toate variantele testului Levene.

Scanned with OKEN Scanner

Test of Homogeneity of Variances

		Levene Statistic	df1	df2	Sig.
TIMP_VINDECARE	Based on Mean	.560	2	60	.574
	Based on Median	.549	2	60	.580
	Based on Median and with adjusted df	.549	2	58.032	.580
	Based on trimmed mean	.531	2	60	.591

Întrucât, în toate variantele sale (bazate pe utilizarea mediei, medianei sau mediei ajustate) testul Levene generat din SPSS, are asociată o probabilitate p peste nivelul prag de 0,05, considerăm faptul că nu avem suficiente dovezi, furnizate de datele analizate la nivelul eșantionului de lucru, pentru a respinge ipoteza nulă H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2$. Prin urmare, testul Levene nu a indicat o diferență semnificativă între varianțele grupurilor comparate ($p > 0{,}05$), ceea ce sugerează că presupunerea de omogenitate a varianțelor este îndeplinită.

c) *populațiile care stau la baza obținerii eșantioanelor de lucru sunt independente*

Cea de a treia presupunere a modelului, populațiile care stau la baza obținerii eșantioanelor de lucru sunt independente, este verificată prin procedura de realizare a studiului, fiecare subiect participant fiind alocat în mod aleatoriu unui singur grup de tratament, anterior demarării acestuia.

Ulterior evaluării presupunerilor modelului ANOVA unifactorială, considerăm că utilizarea analizei varianței este adecvată pentru studierea diferențelor între mediile celor trei populații.

Ipotezele modelului ANOVA unifactorial:

H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: *cel puțin 2 medii sunt diferite*

Testul statistic utilizat este:

$$F_{calculat} = \frac{Varianța\ dintre\ grupuri}{Varianța\ din\ interiorul\ grupurilor}$$

care, atunci când H_0 este adevărată, urmează o distribuție de tip Fisher cu $k - 1 = 3 - 1 = 2$ grade de libertate la numărător, respectiv $n - k = 63 - 3 = 60$ grade de libertate la numitor.

Considerăm un nivel de semnificație $\alpha = 0{,}05$. Valoarea critică a testului F în acest caz este: $F_{critic} = F_{0.05,\ 2,60} = 3{,}150$. Astfel, vom respinge ipoteza nulă în situația când valoarea calculată a testului statistic este mai mare sau egală cu 3,150 (regiunea roșie din partea dreaptă a figurii 10.6). În practică vom putea citi valoarea din tabelul distribuției Fisher, sau vom utiliza funcția Excel: $= F.INV.RT(0{,}05; 2; 60)$.

Scanned with OKEN Scanner

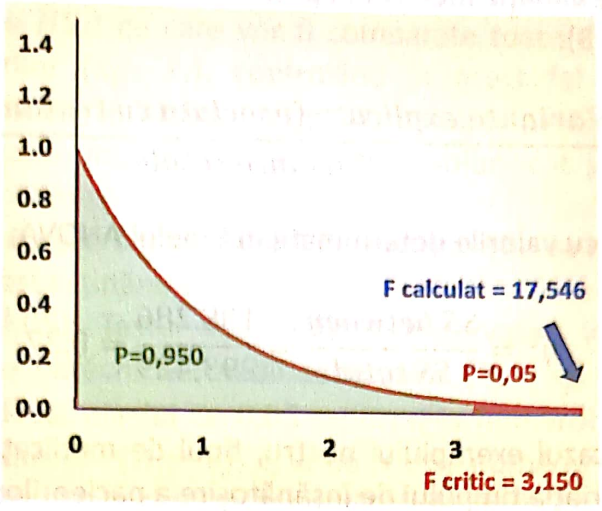


Figura 10.6. Distribuția Fisher cu 2 și 60 grade de libertate

Valoarea calculată $F_{calculat} = F\,(2,60) = 17{,}546$ este mult mai mare decât valoarea critică $F(0.05{,}2{,}60) = 3{,}150$, aspect care evidențiază faptul că există cel puțin două medii ale grupurilor care diferă semnificativ.

ANOVA

TIMP_VINDECARE

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	108.286	2	54.143	17.546	.000
Within Groups	185.143	60	3.086		
Total	293.429	62			

De asemenea, conform tabelului ANOVA realizat în programul SPSS și valoarea probabilității asociate testului este mai redusă comparativ cu nivelul prag de 1% ($p < 0{,}01$), aspect care confirmă respingerea ipotezei nule a egalității mediilor celor trei populații.

În concluzie, *medicamentele 1, 2 și 3 funcționează diferit*, iar diferențele populaționale dintre timpii medii de vindecare, care variază în cazul celor trei tipuri de tratament de la 4 la 7 săptămâni, sunt semnificative statistic. Putem afirma cu un risc de a greși mai mic de 1% faptul că perioada de timp necesară vindecării pacienților diferă în funcție de tratamentul ales. Prin urmare, cele trei tipuri de medicamente studiate au dus la timpi medii de vindecare diferiți: $F(2{,}60) = 17{,}546$ cu $p < 0{,}01$.

Efectul total al tratamentului ne arată ce procent din variația variabilei dependente (timpul necesar fiecărui pacient pentru a se însănătoși) poate fi

Scanned with OKEN Scanner

prezis (explicat) de variația medicației primite (medicamentul 1 – medicamentul 2 – medicamentul 3):

$$\eta^2 = \frac{Varianta\ explicata\ (asociata\ cu\ tratamentul)}{Varianta\ totala}$$

Înlocuind, cu valorile determinate în tabelul ANOVA, obținem:

$$\eta^2 = \frac{SS\ between}{SS\ total} = \frac{108{,}286}{293{,}429} = 0{,}37$$

Astfel, în cazul exemplului nostru, tipul de medicație ales (1, 2 sau 3) explică 37% din variația timpului de însănătoșire a pacienților.

Testul F fiind un test global nu ne permite să evaluăm între care dintre cele trei programe de tratament există diferențe. În consecință, pentru a evidenția diferențele concrete între diferitele perechi de grupuri este necesară utilizarea unor teste de comparații multiple, care să păstreze eroarea de tip I sub nivelul prag de 5%.

10.5. TESTELE DE COMPARAȚII MULTIPLE – ANALIZA POST-HOC

ANOVA reprezintă un test general, care verifică condiția de egalitate a mediilor mai multor populații, fără a specifica însă grupurile concrete între care mediile diferă semnificativ. În acest scop, în situația în care rezultatul testului AVOVA este semnificativ statistic, se vor aplica testele POST-HOC sau testele de comparații multiple, care în esență încearcă să controleze valoarea erorii de tip I α (probabilitatea de a detecta o diferență falsă între mediile populațiilor).

În analiza mai multor perechi de grupuri, probabilitatea de a obține rezultate fals pozitive crește (eroarea de tip I depășește cu mult nivelul de 5% pe măsură ce numărul de grupuri comparate este mai mare). Metodele POST-HOC aplică corecții statistice, cum ar fi cele propuse de Bonferroni, Tukey sau Scheffé, pentru a menține eroarea de tip I (α) la un nivel acceptabil. Acest lucru este esențial în cercetare, pentru a preveni interpretările eronate ale rezultatelor diverselor studii.

Analiza POST-HOC permite o interpretare mai detaliată a rezultatelor ANOVA, facilitând evidențierea mai precisă a relațiilor dintre diferitele perechi de medii ale grupurilor, aspect care conduce în final la formularea unor concluzii utile din punct de vedere practic.

Principalele teste de comparații multiple utilizate în practică, în funcție de designul specific al fiecărui studiu, sunt:

Scanned with OKEN Scanner

1) *Testul Tukey HSD* (*Honest Significant Distance*) utilizează o singură valoare HSD cu care vor fi comparate toate diferențele medii ale grupurilor $(\bar{x}_i - \bar{x}_j)$, controlând în acest fel nivelul de general semnificație α. Astfel, sunt analizate toate perechile de grupuri, atât în situația unor grupuri de același volum cât și pentru grupuri de volum diferit.
2) *Testul Bonferroni* analizează fiecare pereche de medii prin calcularea unui test t, ținând sub control eroarea totală de tip I la nivelul de 5% ($\alpha = 0{,}05$). Testul este foarte conservator, valoarea p aferentă fiecărei perechi de medii comparate $(\bar{x}_i - \bar{x}_j)$ trebuind să fie mai mică decât nivelul de 0,05 raportat la numărul total de comparații efectuate: $N \cdot (N-1)/2$, unde N este numărul de grupuri.
   De exemplu, în cazul comparării unui număr 4 grupuri, vom avea un număr de: $4 \cdot (4-1)/2 = 12/2 = 6$ comparații. Pragul ajustat va fi în acest caz: $0{,}05/6 = 0{,}0083$ și doar diferențele între mediile grupurilor care au o valoare $p < 0{,}0083$ vor fi considerate semnificative statistic. Datorită faptului că este foarte strict, testul Bonferroni poate să nu detecteze în anumite cazuri o diferența reală între grupurile comparate, prezentând astfel un risc mai ridicat de a întâmpina o eroare de tip II (o putere statistică mai scăzută).
3) *Testul Scheffe* compară toate perechile de grupuri, permițând conectarea anumitor grupuri în anumite situații. De exemplu, dacă avem 3 tipuri de tratament (Medicamentul A, Medicamentul B și Placebo) vom putea determina dacă ambele medicamente A și B au un efect diferit față de Palcebo. Cu alte cuvinte se testează grupurile A și B simultan comparativ cu grupul Placebo.
4) *Testul Dunnet* se utilizează atunci când unul dintre grupuri este considerat drept referință (Placebo), iar comparațiile vor fi efectuate între fiecare grup și acesta, folosind o versiune modificată a testului t. Deoarece nu trebuie să compare toate grupurile de medii între ele, testul Dunnet are o putere superioară de a detecta diferențele între grupul de referință și celelalte. Este cel mai indicat test de comparații multiple, în situația studiilor clinice, farmacologice sau experimentale care folosesc un grup de referință.
   Testul Dunnet necesită ca designul studiului să fie echilibrat sau aproape echilibrat (grupurile comparate trebuie să aibă dimensiuni similare).

Revenim la exemplul anterior privind la evaluarea eficienței celor trei tipuri de tratament, în care testul ANOVA prezentat în subcapitolul 10.4 a evidențiat faptul că există cel puțin două medii ale grupurilor care diferă semnificativ. Vom utiliza în cadrul analizei POST-HOC efectuate în programul SPSS testul de comparații multiple *Bonferroni*:

Scanned with OKEN Scanner

Multiple Comparisons

Dependent Variable: TIMP_VINDECARE
Bonferroni

(I) MEDICAMENT	(J) MEDICAMENT	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-2.143*	.542	.001	-3.48	-.81
	3	1.000	.542	.210	-.34	2.34
2	1	2.143*	.542	.001	.81	3.48
	3	3.143*	.542	.000	1.81	4.48
3	1	-1.000	.542	.210	-2.34	.34
	2	-3.143*	.542	.000	-4.48	-1.81

*. The mean difference is significant at the 0.05 level.

Analizând rezultatele testului POST–HOC constatăm că diferența timpilor medii necesari vindecării în cazul tratamentului 1, comparativ cu tratamentul 2 este de $-2{,}14$ săptămâni $(5{,}19 - 7{,}33)$, semnificativă la un $p < 0.01$, fiind situată sub pragul critic de 0,05. Aceasta ne permite să afirmăm cu un risc de a greși mult sub 1 % faptul că subiecții care au urmat tratamentul 2 au nevoie de un timp mai mare pentru a se vindeca comparativ cu cei au urmat tratamentul 1 (figura 10.7), cu aproximativ 2,14 săptămâni (în medie) la nivelul sub-populațiilor generale din care provin aceste grupe.

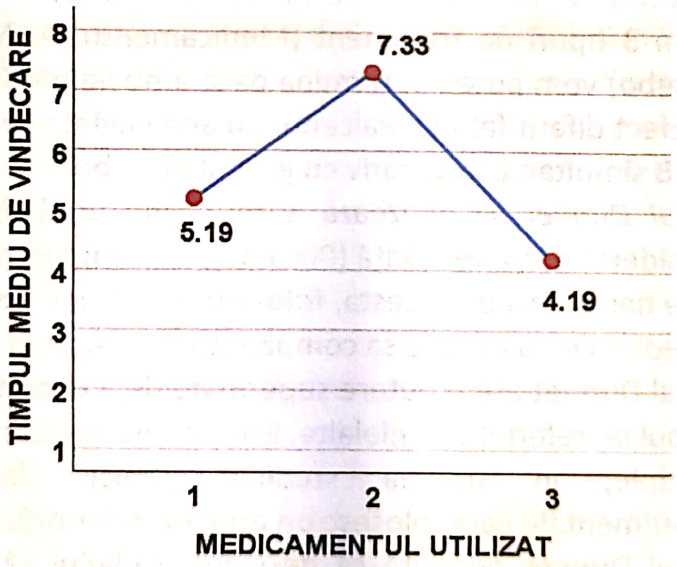


Figura 10.7. Timpii medii de vindecare în cazul celor 3 tratamente

Comparând rezultatele persoanelor implicate în tratamentul 2 comparativ cu cei care au urmat tratamentul 3, diferențele mediilor sunt de 3,14 săptămâni $(7{,}33 - 4{,}19)$ și $p < 0{,}01$. Putem afirma cu un risc de a greși mai mic de 1% că, la nivel populațional, subiecții care au urmat tratamentul 2 au

Scanned with OKEN Scanner

nevoie de un timp mai lung (în medie cu 3,14 săptămâni) comparativ cu subiecții care au urmat tratamentul 3 pentru a se vindeca.

Diferența între timpul timpii medii de vindecare la nivel de eșantion în cazul tratamentelor 1 și 3, de o săptămână în favoarea tratamentului 3, nu este semnificativă statistic ($p = 0{,}21 > 0{,}05$). Prin urmare, la nivelul populațiilor 1 și 3, din care provin aceste două grupuri, nu avem o diferență semnificativă între timpii medii de vindecare, cele două tratamente fiind relativ similare în privința nivelului lor de eficiență. Cu alte cuvinte, diferența de o săptămână în favoarea tratamentului 3 este doar întâmplătoare și nu reflectă influența unor factori esențiali care să îl claseze ca fiind mai eficient în comparație cu tratamentul 2.

Sintetizând, putem afirma că, la nivelul populațiilor generale din care au fost extrase în mod aleatoriu aceste grupuri din eșantion, tratamentul 2 este cel mai puțin eficient, comparativ cu celelalte două, având cel mai îndelungat timp de vindecare, motiv pentru care în cazul replicării experimentului am fi îndreptățiți să renunțăm la acesta, în favoarea celorlalte două.

Scanned with OKEN Scanner

11. CORELAȚIE ȘI REGRESIE SIMPLĂ LINIARĂ

SUBIECTE:

11.1. Corelația simplă liniară – noțiuni preliminare
11.2. Coeficientul de corelație simplă liniară
11.3. Testarea semnificației coeficientului de corelație
11.4. Analiza de regresie simplă liniară - ecuația de regresie și metoda celor mai mici pătrate
11.5. Testarea semnificației pantei dreptei de regresie
11.6. Testarea modelului de regresie simplă liniară prin metoda ANOVA
11.7. Coeficientul de determinare simplă liniară
11.8. Presupunerile modelului de regresie simplă liniară

La finalizarea acestui capitol, studentul va fi capabil să:
O11-1: să explice scopul și importanta coeficientului de corelație
O11-2: să calculeze coeficientul de corelație pentru testarea și interpretarea relației dintre 2 variabile
O11-3: să aplice analiza de regresie simplă liniară pentru a estima legătura dintre 2 variabile
O11-4: să evalueze semnificația pantei ecuației de regresie
O11-5: să utilizeze corect coeficientul de determinare în evaluarea modelului de regresie
O11-6: să evalueze corespunzător condițiile preliminare de aplicare ale modelului de regresie simplă liniară

11.1. CORELAȚIA SIMPLĂ LINIARĂ – NOȚIUNI PRELIMINARE

În cercetarea medicală, înțelegerea relațiilor care pot exista între diferite variabile biologice, clinice sau epidemiologice este esențială pentru formularea corespunzătoare a unor ipoteze statistice privitoare la parametrii unor populații de interes în cadrul diferitelor studii și cercetări cât și pentru luarea unor decizii bine fundamentate din punct de vedere științific.

În contextul analizei științifice medicale moderne, studiul corelației este un pas preliminar important în procesul de explorare a datelor cât și în identificarea posibilelor asocieri semnificative din punct de vedere clinic, care pot

Scanned with OKEN Scanner

sta la baza unor cercetări ulterioare sau chiar la decizii privind efectuarea unor intervenții terapeutice necesare.

Analiza de corelație reprezintă un instrument statistic fundamental care permite investigarea existenței, direcției și intensității unei potențiale relații între două variabile statistice cantitative, numerice (de tip interval sau raport). Pentru a exemplifica, putem vorbi despre corelația existentă între fumat și riscul de a dezvolta cancer pulmonar. Astfel, diferite studii epidemiologice au arătat faptul că există o corelație puternică pozitivă între numărul de țigări fumate pe zi de un pacient și riscul de a dezvolta cancer pulmonar. Prin urmare, cu cât o persoană fumează mai mult, cu atât riscul de cancer pulmonar este mai ridicat, sau, la polul opus, cu cât o persoană fumează mai puțin, riscul de a dezvolta cancer pulmonar este mult diminuat. De asemenea, există o serie de studii de specialitate care au evidențiat o corelație pozitivă între nivelul colesterolului LDL și incidența bolilor cardiovasculare, cum ar fi infarctul miocardic. Un alt exemplu, ar putea fi reprezentat de corelația pozitivă moderată existentă între nivelul vitaminei D și densitatea minerală osoasă, în special în situația pacienților vârstnici.

Creatorul conceptului de corelație a fost antropologul și statisticianul englez Sir Francis Galton (1822-1911), vărul lui Charles Darwin, fondatorul teoriei evoluției speciilor. Fiind interesat de modul în care trăsăturile se transmit de la o generație la alta, în cadrul cercetărilor sale despre ereditate, Galton a observat faptul că, în medie, copiii părinților foarte înalți tind să fie mai scunzi decât părinții lor și, de asemenea, copiii părinților foarte scunzi tind să fie mai înalți decât părinții lor. Aceste rezultate l-au condus în direcția conceptului de *regresie către medie* sau *„regression towards mediocrity"*, Galton concluzionând faptul că, trăsăturile extreme tind să regreseze în direcția mediei populației din generațiile viitoare (Krashniak Adam, Lamm Ehud, Francis Galton's regression towards mediocrity and the stability of types).

Legăturile statistice sunt caracterizate prin faptul că o anumită variabilă X exercită o influență asupra altei variabile Y, îi determină într-o oarecare măsură variația, însă asupra lui Y mai acționează și alți factori în afară de X, care însă, din punctul de vedere al legăturii dintre cele două variabile X și Y, sunt considerați întâmplători.

Legătura statistică dintre două sau mai multe variabile, exprimate sub formă numerică se numește *corelație*. Studiul corelației are la bază într-o primă etapă stabilirea existenței unei legături sau asocieri reale între fenomenele cercetate, urmată apoi, prin utilizarea unor tehnici statistice mai avansate, de caracterizarea direcției, formei și intensității acestei legături.

Sistemul legăturilor statistice dintre fenomenele medicale studiate poate fi clasificat în funcție de mai multe criterii:

a) după numărul variabilelor corelate, deosebim *legături simple* (între două variabile) și *legături multiple* (între trei sau mai multe variabile);

Scanned with OKEN Scanner

b) în funcție de direcția sau sensul lor, legăturile statistice pot fi: *directe* sau pozitive (variabilele variază în același sens) și *inverse* sau negative (variațiile variabilelor au loc în sens contrar);

c) după forma lor, legăturile dintre variabilele statistice pot fi: *liniare* (îmbracă forma unei drepte) sau curbilinii (parabolă, exponențială, hiperbolă, etc.).

În cadrul oricărui studiu statistic asupra corelației dintre variabile trebuie să ținem cont de faptul că analiza este realizată asupra unui număr limitat de observații (eșantionul de lucru), iar pe baza rezultatelor prelucrării acestora se urmărește caracterizarea globală a fenomenului de interes, la nivelul populației generale din care unitățile statistice au fost extrase în mod aleatoriu. Prin urmare, rezultatele analizei de corelație vor avea la bază, în final, o interpretare probabilistică.

Așa cum am prezentat anterior, primul pas în analiza de corelație dintre 2 variabile este reprezentat de realizarea diagramei tip „nor de puncte", pe baza unui sistem de coordonate rectangular. Un anumit punct este obținut la intersecția variantei x_i a variabilei X, reprezentată pe axa orizontală a absciselor, cu varianta corespunzătoare y_i a variabilei Y, reprezentată pe axa verticală a ordonatelor. Totalitatea punctelor de coordonate (x_i, y_i) vor fi prezentate grafic sub forma unui „nor de puncte", care va evidenția în mod sugestiv direcția precum și intensitatea legăturii dintre cele 2 variabile. Această procedură oferă o reprezentare vizuală a relației dintre variabilele studiate.

Corelația dintre două variabile statistice cantitative implică o relație bilaterală între acestea, situându-le astfel la același nivel, fără a face distincție între variabila dependentă și cea independentă din cadrul analizei.

Exemplu: corticosteroizii sunt un grup de hormoni steroizi produși în mod natural de cortexul glandei suprarenale și sintetizați și în formă medicamentoasă pentru utilizare terapeutică. Acești hormoni au efecte puternice asupra metabolismului, inflamației, echilibrului hidro-electrolitic și funcției sistemului imunitar. Corticosteroizii sunt frecvent utilizați în tratamentul unei game largi de afecțiuni inflamatorii și autoimune. Cu toate acestea, utilizarea lor este asociată cu multiple efecte adverse, printre care se numără și hiperglicemia indusă de corticosteroizi. Dorim să verificăm existența unei posibile relații sau asocieri între durata administrării tratamentului cu corticosteroizi, exprimată în zile, și nivelul glicemiei à jeun (mg/dL) în cazul unui eșantion de 50 de pacienți adulți, extras în mod aleatoriu dintr-o populație de nediabetici (tabelul 11.1).

În urma măsurătorilor fiecărui subiect participant privind nivelul glicemiei à jeun, exprimată în mg/dL (variabila Y) și a duratei administrării tratamentului cu corticosteroizi, exprimată în zile (variabila X), suntem interesați să aflăm dacă există o relație sau o legătură între aceste două variabile și în caz afirmativ să realizăm ulterior o estimare a valorilor variabilei Y în baza cunoștințelor pe care le deținem despre nivelurile variabilei X.

Scanned with OKEN Scanner

Tabelul 11.1. Durata tratamentului (variabila X) și nivelul glicemiei à jeun (variabila Y) în cazul eșantionului de 50 de pacienți

Pacient	Variabila X (durata tratamentului - zile)	Variabila Y (glicemia à jeun - mg/dL)
1	67	118
2	18	75
3	84	133
4	58	116
5	87	128
6	48	99
...	...	...
48	71	139
49	27	95
50	81	132

Pentru a contura o primă imagine asupra potențialei relații dintre cele două variabile vom construi *diagrama tip „nor de puncte”* (figura 11.1).

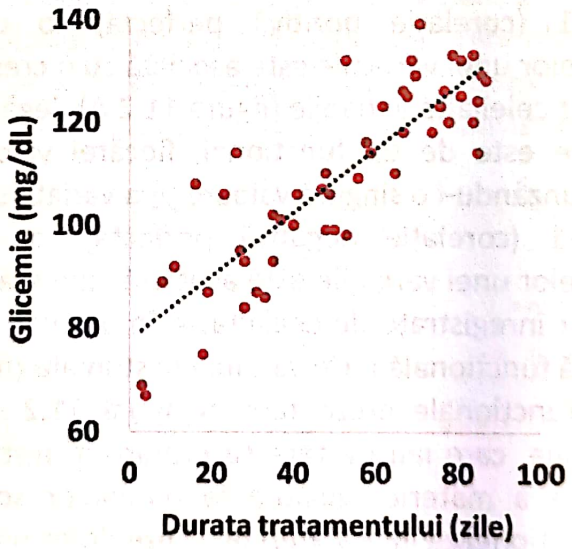


Figura 11.1. Diagrama tip „nor de puncte” pentru variabilele X (durata tratamentului) și Y (nivelul glicemiei à jeun)

Analizând graficul tip „nor de puncte” se observă faptul că cele 50 de puncte (x_i, y_i), aferente pacienților din eșantion, par să fie grupate în jurul unei „drepte invizibile” și, în același timp, faptul că subiecții cu care au urmat un tratament mai îndelungat cu corticosteroizi (variabila X) au de asemenea valori mai ridicate și în cazul glicemiei à jeun (variabila Y).

Scanned with OKEN Scanner

În concluzie, metoda grafică permite evaluarea într-o formă geometrică, deosebit de expresivă a relației dintre două variabile continue (de tip interval sau raport). În funcție de poziția „norului de puncte" în plan poate fi apreciată existența, forma, direcția și intensitatea legăturii dintre variabile. Atunci când punctele se concentrează în jurul unei anumite drepte sau curbe de un anumit tip, acest lucru va sugera existența unei legături care îmbracă această formă geometrică. În același timp, direcția legăturii dintre variabilele studiate poate fi apreciată vizual în funcție de modul de orientare a dreptei sau curbei, care aproximează forma legăturii.

11.2. COEFICIENTUL DE CORELAȚIE SIMPLĂ LINIARĂ

Următorul pas în cadrul analizei corelației, prin care se poate cuantifica intensitatea legăturii liniare dintre două variabile numerice, este reprezentat de determinarea coeficientului de *corelație simplă liniară*. Dezvoltat de matematicianul și statisticianul englez Karl Pearson (1857-1936), coeficientul de corelație, simbolizat prin litera ρ la nivelul populației generale, descrie direcția și intensitatea legăturii dintre 2 variabile numerice aleatoare X și Y, cu repartiții normale $N(\mu_x, \sigma_x^2)$ și $N(\mu_y, \sigma_y^2)$. Acesta poate lua valori cuprinse între:

- $\rho = +1$ (corelație pozitivă perfectă): o creștere a valorilor variantelor unei variabile este asociată cu o creștere proporțională a valorilor celeilalte variabile (figura 11.2 A); legătura dintre cele două variabile este de tip funcțional, fiecărei valori x_i a variabilei X corespunzându-i o singură valoare y_i a variabilei Y și numai una;
- $\rho = -1$ (corelație negativă perfectă): o creștere a valorilor variantelor unei variabile este asociată cu o scădere proporțională a valorilor înregistrate de cealaltă; și în această situație vorbim de o legătură funcțională între variabilele studiate (figura 11.2 B).

Legăturile funcționale prezentate în figura 11.2 se întâlnesc în cazul fenomenelor naturale, care iau naștere ca urmare a mișcării mecanice, fizice, chimice și biologice a materiei. Asupra fenomenelor sociale și a celor din domeniul medical acționează însă alături de cauze determinate, pentru legătura cercetată și o multitudine de alți factori întâmplători, accidentali, a căror acțiune poate influența tendința lor de dezvoltare. Din acest considerent, natura legăturii dintre fenomene, în cazul studiilor statistice medicale, se stabilește în urma unei analize calitative multicriteriale.

Scanned with OKEN Scanner

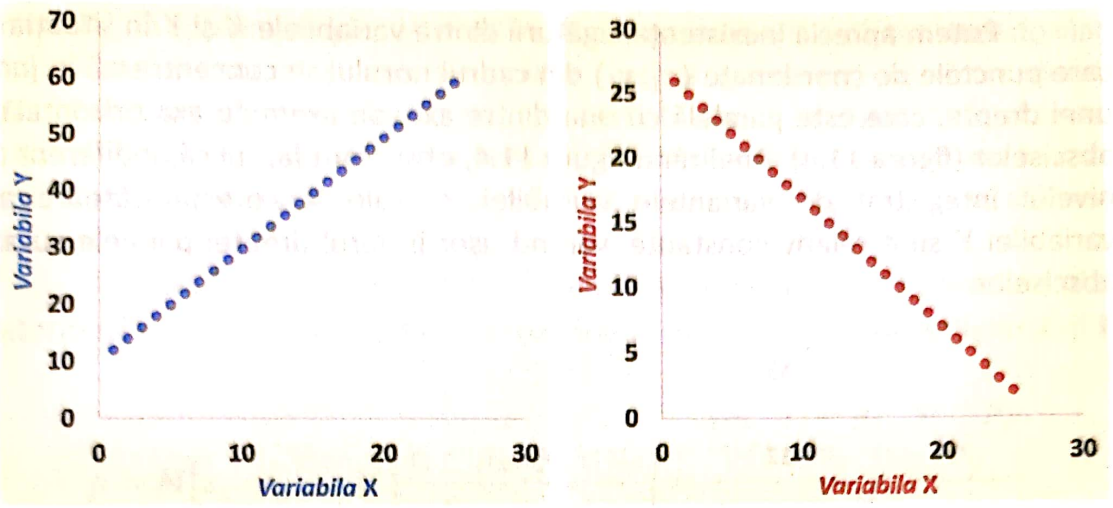


Figura 11.2. Corelație pozitivă perfectă (A) și corelație negativă perfectă (B)

În situația când $\rho = 0$, considerăm faptul că nu există o relație liniară între variabilele studiate. Acest lucru nu înseamnă că cele două variabile sunt independente în general, ci doar necorelate liniar. Așa cum se observă în figura 11.3 B, între variabilele X și Y, deși nu avem o corelație liniară, avem o legătură curbilinie, de tip parabolic foarte puternică. Coeficientul de corelație fiind un indicator al intensității asocierii numai în cazul legăturilor liniare, în situația când între cele 2 variabile statistice există o legătură de tip neliniar, acesta va înregistra o valoare foarte apropiată de zero, similar cu cazul inexistenței vreunei relații.

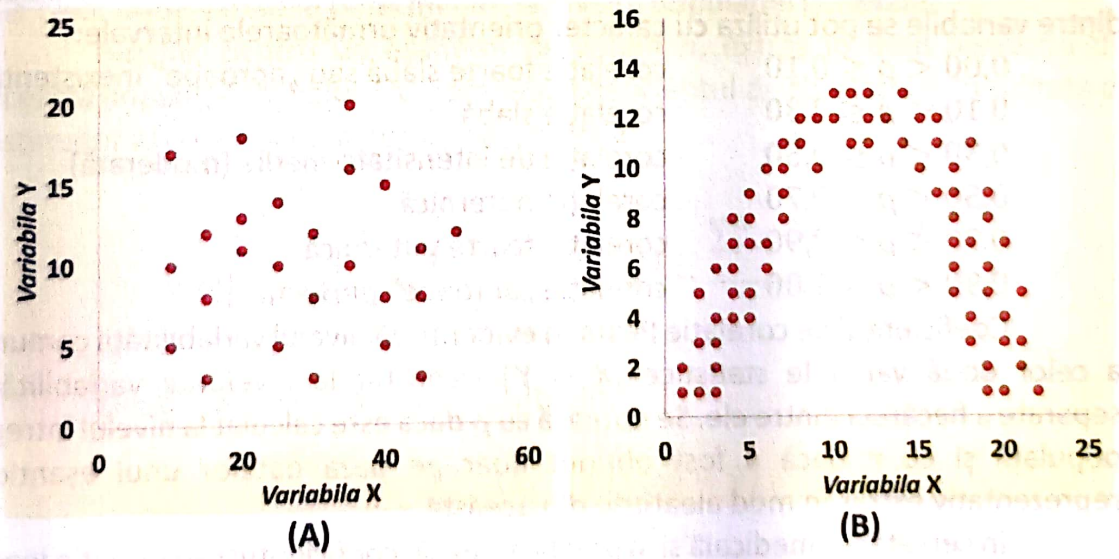


Figura 11.3. Absența corelației / corelației liniare dintre variabilele X și Y

Scanned with OKEN Scanner

Putem aprecia inexistența legăturii dintre variabilele X și Y în situația în care punctele de coordonate (x_i, y_i) din cadrul norului se concentrează în jurul unei drepte, care este paralelă cu una dintre axe, de exemplu axa orizontală a absciselor (figura 11.4). Analizând figura 11.4, observăm faptul că, indiferent de nivelul înregistrat de variantele variabilei X, valorile corespunzătoare ale variabilei Y sunt relativ constante, variind ușor în jurul dreptei paralele cu axa absciselor.

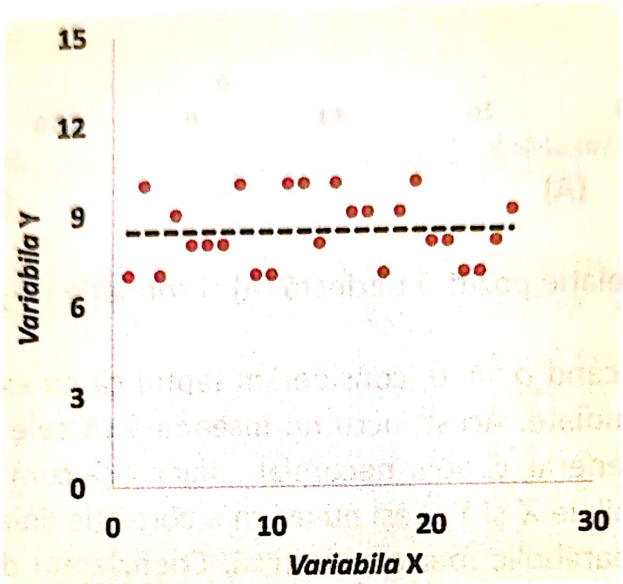


Figura 11.4. Inexistența corelației dintre variabilele X și Y

În analizele practice, pentru a cuantifica intensitatea corelației liniare dintre variabile se pot utiliza cu caracter orientativ următoarele intervale:

$0{,}00 < \rho \leq 0{,}10$	corelație foarte slabă sau „aproape" inexistentă
$0{,}10 < \rho \leq 0{,}30$	corelație slabă
$0{,}30 < \rho \leq 0{,}50$	corelație de intensitate medie (moderată)
$0{,}50 < \rho \leq 0{,}70$	corelație puternică
$0{,}70 < \rho \leq 0{,}90$	corelație foarte puternică
$0{,}90 < \rho \leq 1{,}00$	corelație „aproape" perfectă

Coeficientul de corelație Pearson evidențiază nivelul variabilității comune a celor două variabile statistice (X și Y), raportat la nivelurile variabilității separate a fiecăreia dintre ele. Se notează cu ρ dacă este calculat la nivelul întregii populații și cu r dacă a fost obținut doar pe baza datelor unui eșantion reprezentativ extras în mod aleatoriu din aceasta.

În cercetarea medicală și în practica clinică, coeficientul de corelație joacă un rol deosebit de important, în principal în descoperirea și înțelegerea relațiilor sau asocierilor care pot exista între diferiți factori biologici, comportamentali sau de tratament. În cercetarea clinică de exemplu, permite explorarea legăturii dintre potențiali factori de risc (IMC) și anumiți parametrii biologici (tensiunea

Scanned with OKEN Scanner

arterială); în epidemiologie, este util în studierea asocierilor dintre stilul de viață (persoane fumătoare) și apariția bolilor (cancer pulmonar) sau între durata expunerii la un anumit factor de risc și severitatea bolii; în cazul unui tratament medicamentos permite stabilirea unei relații între nivelul dozei utilizate în cazul unui anumit medicament și răspunsul clinic sau prezența unor efecte adverse, etc.

Coeficientul de corelație ρ la nivelul populației generale de volum N se determină ca o medie aritmetică a produsului scorurilor z ale variabilelor X și Y:

$$\rho = M[z_x \cdot z_y] = M\left[\frac{x_i - \mu_x}{\sigma_x} \cdot \frac{y_i - \mu_y}{\sigma_y}\right] = \frac{\sum_{i=1}^{N} \frac{x_i - \mu_x}{\sigma_x} \cdot \frac{y_i - \mu_y}{\sigma_y}}{N}$$

Efectuând calculele, se obține imediat valoarea lui ρ:

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \mu_x) \cdot (y_i - \mu_y)}{N \cdot \sigma_x \cdot \sigma_y} \tag{11.1}$$

în care:

$\mu_x = \frac{\sum_{i=1}^{n} x_i}{N}$ și $\mu_y = \frac{\sum_{i=1}^{n} y_i}{N}$ reprezintă mediile aritmetice ale celor două variabile la nivelul populației generale, din care provine eșantionul de lucru;

$\sigma_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu_x)^2}{N}}$ și $\sigma_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \mu_y)^2}{N}}$ reprezintă deviațiile standard ale celor două variabile determinate la nivelul populației generale.

Similar, pentru un eșantion de volum n, extras în mod aleatoriu din această populație de volum N, vom avea coeficientul de corelație r, în calitate de estimator al corelației din populația generală:

$$r = M[z_x \cdot z_y] = M\left[\frac{x_i - \bar{x}}{s_x} \cdot \frac{y_i - \bar{y}}{s_y}\right] = \frac{\sum_{i=1}^{N} \frac{x_i - \bar{x}}{s_x} \cdot \frac{y_i - \bar{y}}{s_y}}{n - 1}$$

După realizarea calculelor, obținem:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{(n-1) \cdot s_x \cdot s_y} \tag{11.2.}$$

unde:

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ și $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ sunt mediile aritmetice ale celor două variabile calculate pentru eșantion;

Scanned with OKEN Scanner

$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$ și $s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}$ reprezintă estimatorii deviațiilor standard ale populației generale, determinate în baza datelor din eșantion.

Exemplu: vom prezenta o detaliere a calculelor necesare determinării valorii coeficientului de corelație la nivel de eșantion, pentru a contura o imagine mai clară asupra relației sau legăturii dintre nivelul glicemiei à jeun (variabila Y) și durata administrării tratamentului cu corticosteroizi (variabila X), utilizând datele prezentate în exemplul anterior, pe eșantionul de 50 de subiecți.

Tabelul. 11.2. Calculele necesare pentru determinarea coeficientului de corelație la nivel de eșantion

Pacient	Var. X (x_i)	Var. Y (y_i)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
1	67	118	15,40	8,58	132,13
2	18	75	-33,60	-34,42	1156,51
3	84	133	32,40	23,58	763,99
4	58	116	6,40	6,58	42,11
5	87	128	35,40	18,58	657,73
...	...	...	...	...	...
Total	**2580**	**5471**	**0**	**0**	**18754,40**

Începem prin a determina mediile și deviațiile standard ale celor 2 variabile X și Y, la nivelul datelor cuprinse în eșantion. Se pot utiliza pentru fiecare variabilă formulele Excel $= AVERAGE(valori\ variabila)$ și $= STDEV.S(valori\ variabila)$ sau relațiile standard ale indicatorilor:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2580}{50} = 51{,}60\ zile$$

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{5471}{50} = 109{,}42\ mg/dL$$

$$s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = 24{,}96\ zile$$

$$s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} = 17{,}85\ mg/dL$$

Scanned with OKEN Scanner

Înlocuim mediile și deviațiile standard determinate anterior în formula de calcul 11.2 a coeficientului de corelație:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{(n-1) \cdot s_x \cdot s_y} = \frac{18754{,}40}{(50-1) \cdot 24{,}96 \cdot 17{,}85} = 0{,}86$$

Același rezultat pentru coeficientul de corelație îl putem obține prin aplicarea formulei Excel: $= CORREL(valori\ variabila\ X, valori\ variabila\ Y)$.

Cele 2 variabile sunt pozitiv corelate, în situația în care durata tratamentului (variabila X) înregistrează valori superioare mediei de 51,60 zile, iar valorile pereche ale nivelului glicemiei à jeun (variabila Y) sunt și ele peste valoarea medie de 109,42 mg/dL (punctele din *cadranul I*) și, la polul opus, dacă ambele variabile înregistrează valori perechi sub medie (punctele din *cadranul III*). Astfel, 47 din cele 50 de perechi de valori (x_i, y_i) se situează în cadranele I și III și doar 3 în cadranele II și IV. În cazul celor 47 de perechi de puncte (x_i, y_i), produsele $(x_i - \bar{x}) \cdot (y_i - \bar{y})$ vor fi pozitive și doar pentru 3 perechi de puncte vor fi negative. Prin urmare, suma acestor produse $\sum_{i=1}^{50}(x_i - \bar{x}) \cdot (y_i - \bar{y})$, care apare la numitorul formulei 11.2 în cazul coeficientului de corelație, va fi în mod evident tot pozitivă. Numitorul relației de calcul a coeficientului de corelație, $(n-1) \cdot s_x \cdot s_y$, este întotdeauna un număr pozitiv. În consecință, valoarea calculată a coeficientului de corelație r dintre nivelul glicemiei à jeun și durata administrării tratamentului cu corticosteroizi, determinat la nivelul eșantionului de lucru, va fi în mod evident, în cazul exemplului nostru, pozitivă.

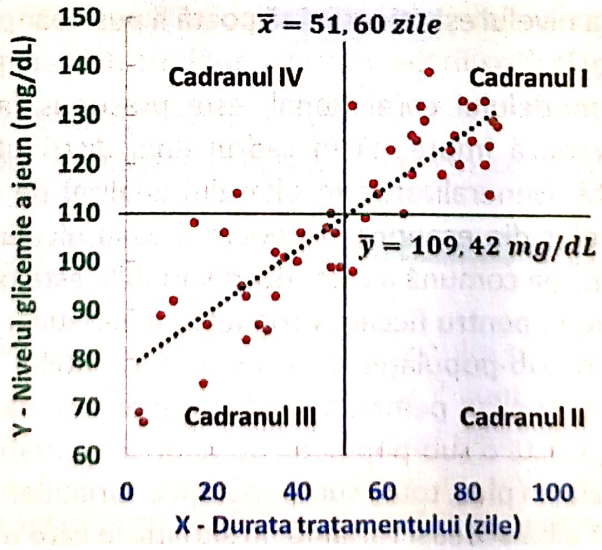


Figura 11.5. Corelația dintre variabilele X (durata tratamentului) și Y (nivelul glicemiei à jeun)

Scanned with OKEN Scanner

Concluzionând, putem afirma faptul că valoarea coeficientului de corelație, calculată la nivel de eșantion $r = 0{,}86$, se situează în intervalul $[0{,}70; 0{,}90]$ și prin urmare la nivelul lotului de studiu, există o relație directă (pozitivă) foarte puternică între cele două variabile. Cu alte cuvinte, pacienții care au urmat un tratament de durată mai lungă cu corticosteroizi au de asemenea valori mai ridicate și în cazul glicemiei à jeun, iar pacienții care au fost tratați o perioadă mai redusă de timp au niveluri mai scăzute și în cazul glicemiei à jeun.

11.3. TESTAREA SEMNIFICAȚIEI COEFICIENTULUI DE CORELAȚIE

Este de interes în cadrul unei anumite cercetări să putem cunoaște dacă $\rho \neq 0$, sau cele două variabile X și Y sunt corelate liniar la nivel populațional. Deoarece coeficientul de corelație de la nivelul populației ρ este în general necunoscut, vom extrage în mod aleatoriu un eșantion din aceasta și vom calcula coeficientul de corelație r din lotul studiat, cu titlul de estimator punctual al lui ρ. În consecință, vom testa ipoteza nulă a absenței corelației în populația generală $H_0: \rho = 0$ împotriva ipotezei alternative a existenței corelației la nivel populațional $H_1: \rho \neq 0$. Dorim astfel să cunoaștem dacă valoarea obținută pentru coeficientul r la nivel de eșantion este suficient de mare încât să evidențieze și o corelație la nivel populațional. Cu alte cuvinte, se pune următoarea întrebare: există posibilitatea ca valoarea coeficientului de corelație r, calculată pe baza datelor din eșantion, să provină dintr-o populație de observații pereche (x_i, y_i) în care corelația să fie de fapt absentă și în consecință rezultatul obținut la nivelul eșantionului să poată fi pus doar pe seama întâmplării sau erorii de sondaj?

În cadrul modelului corelațional, este presupus faptul că, cele două variabile X și Y variază împreună în cadrul unei distribuții comune, numită distribuție bivariată. Generalizarea rezultatului analizei de corelație, efectuată inițial pe baza datelor din eșantion, se poate face la nivelul populației doar în situația când distribuția comună a celor două variabile este o distribuție normală bivariată. Prin urmare, pentru fiecare variantă x_i a variabilei X, care este normal distribuită, există o sub-populație de valori ale variabilei Y, care urmează o distribuție normală. Similar, pentru fiecare variantă y_i a variabilei Y, care este normal distribuită, există o sub-populație de valori ale variabilei X, care urmează o distribuție normală. În plus, toate sub-populațiile variabilelor X și Y, evidențiate anterior, trebuie să aibă aceeași varianță. În situația în care distribuția bivariată a variabilelor X și Y nu poate fi considerată ca fiind normală sau are o formă necunoscută, analiza statistică inferențială va conduce la rezultate necorespunzătoare, valoarea determinată a coeficientului de corelație fiind valabilă doar pentru studiul datelor de la nivelul eșantionului investigat.

Scanned with OKEN Scanner

Rezultă imediat faptul că, în demararea analizei de corelație simplă liniară dintre două variabile statistice continue (de tip interval sau raport) este necesar să verificăm pe lângă normalitatea distribuției fiecărei variabile în parte și normalitatea distribuției bivariate a acestora. În acest sens, în practică pot fi aplicate testele Shapiro-Wilk generalizat (o extensie a testului Shapiro-Wilk univariat) sau Mardia (care verifică asimetria și boltirea multivariate). Atunci când lucrăm cu eșantioane de volum mare ($n \geq 30$ de unități statistice), verificarea normalității fiecărei variabile nu mai este strict necesară, datorită teoremei limită centrală. De asemenea, trebuie ținut cont de faptul că fiecare pereche de observații (x_i, y_i) trebuie să aparțină doar unei singure unități statistice studiate (să fie independentă față de celelalte).

Testul statistic bilateral utilizat pentru a testa ipoteza nulă $H_0: \rho = 0$ împotriva ipotezei alternative $H_1: \rho \neq 0$, este:

$$t_{calculat} = r \cdot \sqrt{\frac{n-2}{1-r^2}} \tag{11.3}$$

care, atunci când H_0 este adevărată și presupunerile modelului se consideră a fi îndeplinite, urmează o distribuție Student (t) cu $n-2$ grade de libertate.

În cadrul relației 11.3, r reprezintă coeficientul de corelație calculat la nivelul eșantionului de lucru, iar n reprezintă volumul acestuia sau numărul de observații pereche (x_i, y_i).

Nivelul de semnificație $\alpha = 0{,}05$ va fi împărțit în mod egal între cele două cozi laterale $\alpha/2 = 0{,}025$ asociat cu valorile foarte mici și $\alpha/2 = 0{,}025$ asociat cu valorile foarte mari. Pentru determinarea valorii critice $t_{n-2,\,\frac{\alpha}{2}}$ putem folosi tabelele distribuției Student.

Regiunea de respingere a testului t va fi în consecință compusă din 2 părți:

$$t_{calculat} \leq -t_{n-2,\,\frac{\alpha}{2}} \quad \text{și} \quad t_{calculat} \geq t_{n-2,\,\frac{\alpha}{2}}$$

În situația în care valoarea calculată a testului se situează în regiunea de respingere, vom concluziona că variabilele X și Y sunt corelate liniar în cadrul populației generale din care a fost extras în mod aleatoriu eșantionul de lucru (respingem ipoteza nulă a absenței corelației în populația generală). Dacă valoarea calculată a testului nu se află poziționată în regiunea de respingere, vom afirma că nu există suficiente dovezi, furnizate de eșantionul studiat, pentru a putea concluziona faptul că variabilele sunt corelate și la nivel populațional. De asemenea, în funcție de valoarea probabilității p aferentă testului utilizat vom decide dacă:

Scanned with OKEN Scanner

- ✓ în cazul în care $p < 0{,}05$ (pentru un nivel de semnificație $\alpha = 5\%$), respingem ipoteza nulă H_0 a absenței corelației în populația generală;
- ✓ în cazul în care $p \geq 0{,}05$ (pentru un nivel de semnificație $\alpha = 5\%$), nu respingem ipoteza nulă H_0 și considerăm că, în populația din care a fost extras în mod aleatoriu eșantionul studiat corelația ar putea fi absentă.

Exemplu: folosind datele din exemplul anterior, vom testa semnificația coeficientului de corelație ρ al populației generale prin aplicarea testului Student, folosind ca estimator punctual coeficientul de corelație $r = 0{,}86$ determinat pe baza datelor din eșantionul de volum $n = 50$ pacienți studiat anterior.

Pentru a testa ipoteza nulă $H_0: \rho = 0$ împotriva ipotezei alternative ($H_1: \rho \neq 0$), vom utiliza relația 11.3:

$$t_{calculat} = r \cdot \sqrt{\frac{n-2}{1-r^2}} = 0{,}86 \cdot \sqrt{\frac{50-2}{1-(0{,}86)^2}} = 11{,}63$$

Valoarea critică o citim din tabelul distribuției Student sau o putem obține direct în programul Excel, cu formula: $= T.INV(0{,}975; 48)$:

$$t_{critic} = t_{48;0{,}025} = 2{,}01$$

Având în vedere faptul că $t_{calculat} > t_{critic}$ (ne situăm în partea dreaptă a regiunii de respingere), respingem ipoteza nulă H_0 a absenței corelației din populația generală. Cu alte cuvinte coeficientul de corelație calculat la nivel de eșantion ($r = 0{,}86$) este prea mare ca să fi putut proveni dintr-o populație în care corelația să fost zero. Astfel, corelația prezentă și evidențiată la nivel de eșantion nu este rezultatul întâmplării sau hazardului ci este un rezultat semnificativ statistic.

În același timp, utilizând programul SPSS pentru analiza aceluiași set de date, observăm faptul că și valoarea probabilității p asociate testului statistic utilizat sau nivelul de semnificație al acestuia (Sig. 2-tailed) este sub pragul de 5% ($p = 0{,}000 < 0{,}05$) și chiar sub cel de 1% ($p < 0{,}01$).

Correlations

		Durata_tratament_zile	Glicemie_mg_dL
Durata_tratament_zile	Pearson Correlation	1	.859**
	Sig. (2-tailed)		.000
	N	50	50
Glicemie_mg_dL	Pearson Correlation	.859**	1
	Sig. (2-tailed)	.000	
	N	50	50

**. Correlation is significant at the 0.01 level (2-tailed).

Scanned with OKEN Scanner

Prin urmare, considerăm că avem suficiente dovezi în sprijinul respingerii ipotezei nule H_0, iar corelația observată la nivelul lotului de studiu ($r = 0{,}86$) este semnificativă statistic la nivel populațional.

Rezultatele testului t pentru semnificația coeficientului de corelație Pearson indică o asociere pozitivă puternică la nivel populațional între cele două variabile studiate, respectiv nivelul glicemiei à jeun și durata administrării tratamentului cu corticosteroizi ($r = 0{,}86;\ t(48) = 11{,}63;\ p < 0{,}01$).

11.4. ANALIZA DE REGRESIE SIMPLĂ LINIARĂ - ECUAȚIA DE REGRESIE ȘI METODA CELOR MAI MICI PĂTRATE

Analiza de regresie simplă liniară reprezintă o procedură statistică de estimare a unei legături dintre o variabilă dependentă (prezisă sau rezultat) și o variabilă independentă (predictor sau variabilă explicativă), sintetizată sub forma unei relații, care are drept scop estimarea valorilor variabilei dependente Y, utilizând valorile variabilei independente X. În cadrul analizei, ambele variabile trebuie să fie numerice (de tip interval sau raport).

Ecuația de regresie reprezintă forma matematică care exprimă relația dintre variabile. Cazul cel mai simplu de legătură dintre două variabile X și Y este cel al regresiei liniare, când mulțimea punctelor de coordonate (x_i, y_i) se distribuie geometric în jurul unei drepte:

$$y_i = \hat{y}_i + \varepsilon_i = a + b \cdot x_i + \varepsilon_i \tag{11.4}$$

în care:

- y_i sunt valorile observate la nivelul eșantionului de studiu ale variabilei Y, determinate în urma finalizării experimentului (ordonatele punctelor din nor);
- $\hat{y}_i = a + b \cdot x_i$ reprezintă valorile estimate ale variabilei Y, utilizând formula dreptei de regresie (ordonatele punctelor corespunzătoare celor observate, care sunt situate pe dreapta de regresie);
- $\varepsilon_i = y_i - \hat{y}_i$ componenta de eroare a modelului, bazată pe valorile reziduale (distanțele verticale dintre valorile observate și cele corespunzătoare estimate – figura 11.6);
- a = ordonata la origine (valoarea estimată a lui Y atunci când $x_i = 0$, sau punctul în care dreapta de regresie intersectează axa ordonatelor);
- b = coeficientul de regresie, care măsoară înclinația, panta dreptei de regresie față de axa absciselor ($b = tg\alpha$) sau variația variabilei dependente Y la fiecare modificare cu o unitate a lui X. Astfel, dacă:

$$x' = x + 1 \implies y' = a + b \cdot (x + 1) = a + b \cdot x + b \implies y' = y + b$$

Scanned with OKEN Scanner

Metoda de calcul pentru parametrii dreptei de regresie, care permite alegerea celei mai bune drepte, constă în a minimiza eroarea modelului. Deoarece o parte dintre nivelurile erorii ε_i sunt pozitive, iar altele sunt negative, pentru a nu pierde din informație, se va utiliza în calcul pătratul erorii (figura 11.6).

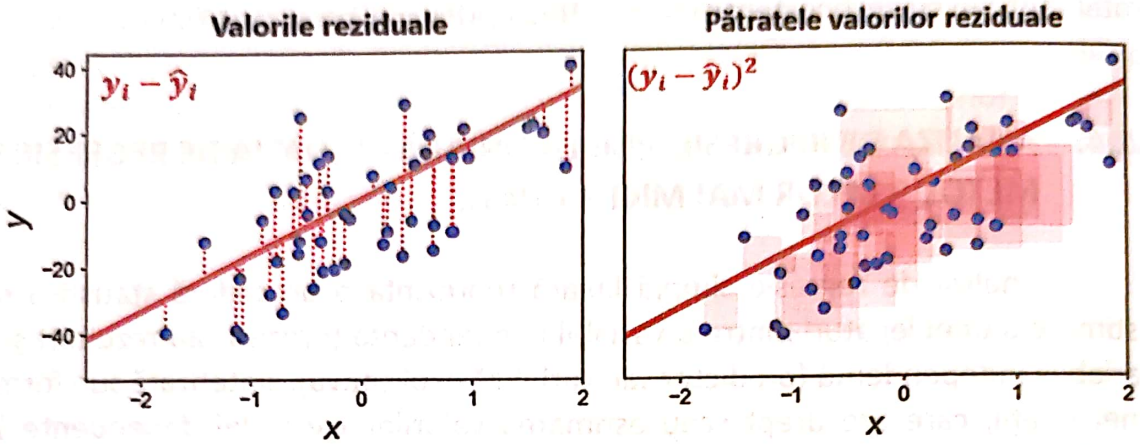


Figura 11.6. Valorile reziduale $\varepsilon_i = y_i - \hat{y}_i$ și pătratele acestora $\varepsilon_i^2 = (y_i - \hat{y}_i)^2$ în cazul variabilei dependente Y

Metoda celor mai mici pătrate permite alegerea *„celei mai bune drepte de regresie"*, astfel încât să minimizeze suma pătratelor distanțelor verticale ε_i^2 dintre valorile observate y_i (din norul de puncte) și cele estimate $\hat{y}_i$ (prezise - situate pe dreapta de regresie) ale variabilei dependente Y (relația 11.5).

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - b \cdot x_i)^2 = minim \tag{11.5}$$

Eroarea totală $\sum_{i=1}^{n} \varepsilon_i^2$ va fi formată astfel din suma tuturor pătratelor erorilor determinate de cele n valori experimentale. Pentru a minimiza valoarea erorii totale este necesar ca derivatele parțiale în funcție de fiecare variabilă (a și b) din ecuația de regresie 11.4 să fie egale cu zero.

În urma efectuării calculelor, se obține următorul sistem de ecuații:

$$\begin{cases} \sum_{i=1}^{n} y_i = na + b \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 \end{cases} \tag{11.6}$$

Scanned with OKEN Scanner

Prin rezolvarea sistemului de ecuații prezentat în relațiile 11.6 se vor obține valorile celor doi parametri ai ecuației de regresie a și b.

➢ *Parametrul b (panta dreptei de regresie):*

$$b = r \cdot \frac{s_y}{s_x} \tag{11.7}$$

unde: $s_x = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$ și $s_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}}$ sunt deviațiile standard ale variabilelor X si Y, calculate la nivel de eșantion, iar $r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{(n-1)\cdot s_x \cdot s_y}$ este coeficientul de corelație dintre acestea;

➢ *Parametrul a (intersecția cu axa verticală Oy):*

$$a = \bar{y} - b \cdot \bar{x} \tag{11.8}$$

unde: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ și $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ sunt mediile aritmetice ale variabilelor X si Y, calculate la nivel de eșantion.

Dreapta de regresie obținută prin metoda celor mai mici pătrate, va trece întotdeauna prin punctul de coordonate $(\bar{x}, \bar{y})$. Acest lucru rezultă imediat, prin împărțirea primei ecuații din sistemul 11.6 la n:

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i \quad \Big| : n \quad \Rightarrow \quad \frac{\sum_{i=1}^{n} y_i}{n} = a + b \cdot \frac{\sum_{i=1}^{n} x_i}{n}$$

sau: $\bar{y} = a + b \cdot \bar{x}$ (11.9)

În relația 11.9 observăm faptul că: $\frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$ și $\frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}$ sunt mediile celor două variabile X si Y. Rezultă imediat $\bar{y} = a + b \cdot \bar{x}$ sau, cu alte cuvinte, dreapta de regresie $y_i = a + b \cdot x_i$ va trece întotdeauna prin punctul de coordonate $(\bar{x}, \bar{y})$.

Exemplu: continuând pe linia exemplului anterior, vom determina cei doi parametri a și b ai ecuației de regresie, folosind metoda celor mai mici pătrate, și vom realiza și reprezentarea grafică a dreptei de regresie.

Utilizând programul Excel (meniul *„Data"*, modulul „Data Analysis", opțiunea *„Regression"*, vom determina valorile coeficienților de regresie a și b:

	Coefficients
Intercept - a	77,712
Durata tratamentului (zile) - b	0,614

Scanned with OKEN Scanner

Reprezentăm grafic dreapta de regresie (figura 11.7), cu parametrii a și b determinați anterior:

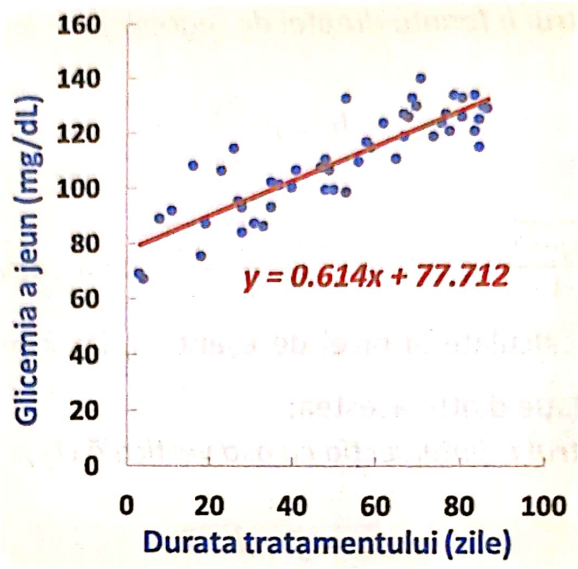


Figura 11.7. Dreapta de regresie $\hat{y}_i$ care minimizează suma tuturor pătratelor erorilor

Dreapta de regresie din figura 11.7 aproximează în modul cel mai fidel forma și direcția legăturii dintre cele două variabile (durata tratamentului cu corticosteroizi și nivelul glicemiei à jeun). Cu alte cuvinte se poate afirma faptul că nu există nicio altă dreaptă, care poate fi trasată prin mijlocul norului de puncte, pentru care suma pătratelor diferențelor dintre valorile observate și cele estimate sau prezise de model să fie mai redusă.

În al doilea rând, așa cum am arătat în relația 11.9, dreapta de regresie trece prin punctul de coordonate $(\bar{x}, \bar{y})$, unde $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{2580}{50} = 51{,}60\ zile$ și $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{5471}{50} = 109{,}42\ mg/dL$ au fost determinate anterior.

Pentru a verifica, introducem în ecuația de regresie valoarea medie a variabilei X (durata tratamentului) $\bar{x} = 51{,}60\ zile$ și vom obține valoarea medie a glicemiei $\bar{y} = 109{,}42\ mg/dL$:

$$\hat{y}_i = a + b \cdot \bar{x} = 77{,}712 + 0{,}614 \cdot 51{,}600 = 109{,}42\ mg/dL = \bar{y}$$

Prin urmare, valoarea estimată a glicemiei în cazul unui tratament cu o durată medie $\bar{x} = 51{,}60\ zile$ este egală cu media valorii glicemiei lotului de pacienți considerat, respectiv $\bar{y} = 109{,}42\ mg/dL$. Rezultă că punctul de coordonate $(\bar{x}, \bar{y}) = (51{,}60\ zile, 109{,}42\ mg/dL)$ va fi situat pe dreapta de regresie construită anterior, după determinarea parametrilor a și b prin utilizarea metodei celor mai mici pătrate.

Scanned with OKEN Scanner

11.5. TESTAREA SEMNIFICAȚIEI PANTEI DREPTEI DE REGRESIE

În cadrul analizei vom aplica testul statistic Student, pentru a verifica dacă panta dreptei de regresie a populației β din care a fost extras eșantionul cu panta b este diferită de zero. În situația respingerii ipotezei nule ($H_0: \beta = 0$) putem utiliza dreapta de regresie $y_i = a + b \cdot x_i$ pentru a prezice valorile variabilei dependente Y în funcție de cea independentă X. Dacă nu avem suficiente dovezi pentru a respinge ipoteza nulă ($H_0: \beta = 0$), rezultă că putem utiliza în analiză media variabilei Y ca predictor în locul ecuației de regresie. Cu alte cuvinte, dreapta de regresie este orizontală ($panta = 0$) și astfel nu există nicio legătură între cele 2 variabile studiate.

Ipotezele nulă și alternativă ale testului statistic sunt:

- $H_0: \beta = 0$ (panta dreptei de regresie a populației generale este zero)
- $H_1: \beta \neq 0$ (panta dreptei de regresie a populației este diferită de zero)

Vom utiliza litera β din alfabetul grecesc pentru a evidenția panta dreptei de regresie din populația generală și cu b pe cea determinată la nivel de eșantion.

Testul statistic utilizat pentru a testa ipoteza nulă $H_0: \beta = 0$ este:

$$t_{calc} = \frac{b - \beta}{s_b} = \frac{b - 0}{s_b} \tag{11.10}$$

care, atunci când H_0 este adevărată, urmează o distribuție Student cu $n - 2$ grade de libertate.

b este estimatorul pantei de regresie, calculat la nivelul eșantionului de volum n, iar s_b este eroarea standard a pantei de regresie, calculată tot pe baza datelor din eșantion:

$$s_b = \sqrt{\frac{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{11.11}$$

Deoarece utilizăm un test bilateral, nivelul de semnificație $\alpha = 0{,}05$ va fi împărțit în mod egal între cele două cozi laterale ale distribuției Student.

În consecință, regiunea de respingere a testului t va fi:

$$t_{calculat} \leq -t_{n-2,\,\frac{\alpha}{2}} \quad \text{și} \quad t_{calculat} \geq t_{n-2,\,\frac{\alpha}{2}}$$

În situația în care valoarea calculată a testului se situează în regiunea de respingere, vom concluziona că panta dreptei de regresie este diferită de zero în populația generală din care a fost extras în mod aleatoriu eșantionul de lucru

Scanned with OKEN Scanner

(respingem ipoteza nulă a pantei dreptei de regresie zero $H_0: \beta = 0$). Dacă valoarea calculată a testului nu se află poziționată în regiunea de respingere, vom afirma că nu există suficiente dovezi, furnizate de eșantionul studiat, pentru a putea concluziona faptul că panta dreptei de regresie este diferită de zero în populația generală. În funcție de valoarea probabilității p aferentă testului statistic utilizat vom decide dacă:

- ✓ în cazul în care $p < 0{,}05$ (pentru un nivel de semnificație de 5%), respingem ipoteza nulă $H_0: \beta = 0$ în populația generală și putem utiliza dreapta de regresie în scopul realizării unor predicții asupra valorilor variabilei dependente în funcție de variabila independentă;
- ✓ în cazul în care $p \geq 0{,}05$ (pentru un nivel de semnificație de 5%), nu respingem ipoteza nulă H_0 și considerăm că, în populația din care a fost extras în mod aleatoriu eșantionul studiat panta dreptei de regresie ar putea fi zero sau, cu alte cuvinte, nu există nicio legătură între cele două variabile studiate.

Exemplu: utilizând în continuare datele prezentate anterior, vom testa dacă putem concluziona faptul că panta dreptei de regresie a populației care descrie relația dintre cele 2 variabile: X (durata tratamentului cu corticosteroizi) și Y (nivelul glicemiei à jeun) este zero $H_0: \beta = 0$.

Utilizând din nou programul Excel (meniul *„Data"*, modulul „Data Analysis", opțiunea *„Regression"*, vom determina valorile coeficienților de regresie a și b, erorile standard aferente, valorile testului t calculat în cazul fiecărui coeficient, valorile probabilităților p asociate acestor teste și intervalele de încredere corespunzătoare:

	Coefficients	*Std. Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	77,712	3,023	25,709	0,000	71,634	83,790
Durată tratament	0,614	0,053	11,630	0,000	0,508	0,721

$$t_{calculat} = \frac{b - 0}{s_b} = \frac{0{,}614 - 0}{0{,}053} = 11{,}630$$

$$t_{critic} = t_{n-2,\ \frac{\alpha}{2}} = t_{(50-2),\ 0.05} = 2{,}011$$

În concluzie, vom respinge ipoteza $H_0: \beta = 0$ deoarece $t_{calculat} > t_{critic}$ concluzionând că panta dreptei „adevărate" de regresie a populației generale nu este zero ($p < 0{,}01$).

Implicația practică a acestor rezultate se referă la faptul că, este de așteptat să obținem predicții sau estimări mai bune ale valorilor variabilei Y dacă utilizăm ecuația de regresie determinată la nivel de eșantion, comparativ cu

Scanned with OKEN Scanner

situația în care am ignora relația dintre variabilele X și Y. Având în vedere că valoarea coeficientului de regresie $b = 0{,}614$ este un număr pozitiv, rezultă ca și relația dintre durata tratamentului cu corticosteroizi și nivelul glicemiei à jeun este o relație directă liniară.

Interpretarea practică a valorii coeficientului de regresie ($b = 0{,}614$): pentru fiecare zi de tratament cu corticosteroizi în plus (variabila X), nivelul estimat al glicemiei à jeun (variabila Y) va crește cu 0,614 mg/dL. Deoarece relația este liniară, creșterea este constantă, respectiv 0,614 mg/dL pentru fiecare zi suplimentară de tratament. De exemplu, dacă un pacient urmează un tratament cu corticosteroizi timp de 30 de zile, ne așteptăm să înregistrăm o creștere a glicemiei à jeun a acestuia cu 18,42 mg/dl ($30 \cdot 0{,}614$ mg/dL) comparativ cu nivelul glicemiei anterior tratamentului.

Ulterior obținerii parametrului b al ecuației de regresie, deseori în practică suntem interesați să determinăm și un interval de încredere estimat cu o probabilitate de 95% pentru parametrul β (panta dreptei de regresie) al populației generale pentru un nivel de semnificație $\alpha = 0{,}05$.

În vederea determinării intervalului de încredere pentru parametrul b al ecuației de regresie, vom utiliza relația clasică, detaliată în cadrul capitolului al 7-lea: Estimarea și intervalele de încredere:

$$estimatorul \pm t_{n-2,\,\frac{\alpha}{2}} \cdot eroarea\ standard\ a\ estimarii \qquad (11.12)$$

în care:

- *estimatorul* este reprezentat de parametrul b, panta dreptei de regresie ($b = 0{,}614$);
- *eroarea standard a estimarii* s_b, care se determină prin aplicarea formulei 11.11. ($s_b = 0{,}053$)

Așa cu se observă din rezultatele obținute prin utilizarea programului Excel, intervalul de încredere 95% pentru panta dreptei de regresie de la nivel populațional este (0,508; 0,721). Prin urmare, putem avea încredere în proporție de 95% că intervalul de confidență determinat anterior pentru panta dreptei de regresie β a populației generale va include valoarea reală și necunoscută a acesteia. Din punct de vedere probabilistic, vom putea afirma faptul că, în cadrul procesului de eșantionare repetată, 95% dintre intervalele de încredere determinate în acest mod vor include valoarea „reală" a pantei dreptei de regresie a populației. În plus, observăm faptul că acest interval de încredere nu conține valoarea zero, fapt care ne arată odată în plus, că această valoare nu reprezintă un nivel probabil sau un candidat potențial pentru parametrul β estimat.

În cazul regresiei simple liniare, cele două teste: pentru verificarea semnificației pantei dreptei de regresie și pentru verificarea semnificației

Scanned with OKEN Scanner

coeficientului de corelație sunt echivalente, ambele folosind o statistică Student t cu $n-2$ grade de libertate. Rezultatele acestor teste sunt identice ($t_{calculat} = 11{,}630$ în cazul exemplului nostru), singura diferență fiind reprezentată doar de modalitatea specifică de interpretare:

- *Testul pentru verificarea semnificației pantei dreptei de regresie* evaluează relația de dependență dintre variabile, respectiv modificarea variabilei dependente Y în funcție de modificarea cu o unitate a variabilei independentei X, având un rol predictiv și o direcție clară;
- *Testul pentru verificarea semnificației coeficientului de corelație* evaluează doar relația de asociere liniară între cele două variabile, neavând valoare predictivă.

11.6. TESTAREA MODELULUI DE REGRESIE SIMPLĂ LINIARĂ PRIN METODA ANOVA

Pentru a evalua performanța modelului de regresie privind legătura dintre cele 2 variabile statistice se utilizează în practică analiza variației (ANOVA). Aceasta se bazează pe legea variației totale, în care varianța observată într-o anumită variabilă este împărțită în componente atribuibile diferitelor surse de variație (figura 11.8). La o simplă analiză vizuală a figurii 11.8 remarcăm o împrăștiere relativ mai mare a punctelor din nor în jurul mediei $\bar{y}$ variabilei Y comparativ cu cea existentă în jurul dreptei de regresie ($\hat{y}_i$).

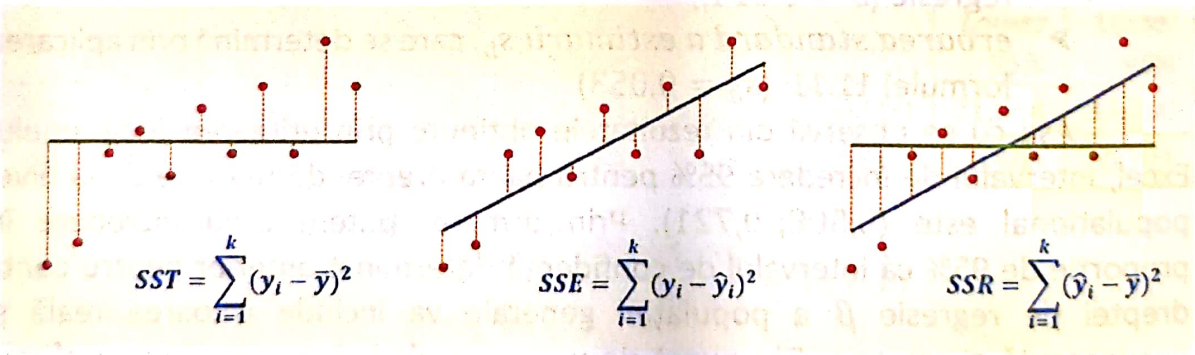


Figura 11.8. Descompunerea variației totale a variabilei dependente Y

Aprecierea vizuală a utilității dreptei de regresie în carul analizei în raport cu valoarea medie a variabilei nu este însă suficientă, fiind necesară descompunerea variației totale în funcție de principalii factori de influență:

a) *variația totală* $(y_i - \bar{y})$ sau abaterea fiecărui nivel observat y_i de la nivelul mediu al variabilei dependente Y; aceasta evidențiază cât de mult

Scanned with OKEN Scanner

variază variantele y_i ale variabilei Y, în general, fiind independentă de model;

b) *variația explicată* $(\hat{y}_i - \bar{y})$ evidențiază distanța verticală a fiecărei valori a variabilei dependente de la dreapta de regresie $(\hat{y}_i)$ la cea orizontală a nivelului mediu $(\bar{y})$ sintetizând acea parte din variația variabilei dependente Y care poate fi explicată de modelul de regresie liniară;

c) *variația neexplicată* $(y_i - \hat{y}_i)$ reprezintă distanța verticală dintre fiecare valoare observată y_i și punctul corespunzător $\hat{y}_i$ situat pe dreapta de regresie sau componenta de eroare a modelului ε_i; aceasta arată cât de departe se situează valorile reale de cele estimate prin intermediul modelului de regresie.

Prin urmare, pentru fiecare diferență dintre valorile y_i ale variabilei dependente Y și nivelul mediu al seriei $\bar{y}$, prin scăderea și adunarea nivelului $\hat{y}_i$ situat pe dreapta de regresie (estimat) vom obține:

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \tag{11.13}$$

Rezultă imediat că pentru fiecare variantă y_i a variabilei dependente Y, se verifică relația: variația totală = variația neexplicată + variația explicată.

Ridicând la pătrat și însumând relația 11.13 se păstrează egalitatea:

$$\sum_{i=1}^{k}(y_i - \bar{y})^2 = \sum_{i=1}^{k}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{k}(\hat{y}_i - \bar{y})^2 \tag{11.14}$$

în care:

Suma totală a pătratelor: $SST = \sum_{i=1}^{k}(y_i - \bar{y})^2$ reprezintă suma pătratelor diferențelor verticale dintre valorile observate ale variabilei dependente și media lor, arătând cât de bine le aproximează valoarea medie a acestora;

Suma pătratelor erorilor: $SSE = \sum_{i=1}^{n}\varepsilon_i^2 = \sum_{i=1}^{k}(y_i - \hat{y}_i)^2$ reprezintă suma pătratelor diferențelor verticale dintre valorile observate ale variabilei dependente și valorile estimate (prezise) prin dreapta de regresie. Această variație reprezintă expresia factorului de eroare. Valoarea SSE este minimizată prin aplicarea metodei celor mai mici pătrate în cadrul analizei modelului de regresie simplă liniară.

Suma pătratelor modelului de regresie: $SSR = \sum_{i=1}^{k}(\hat{y}_i - \bar{y})^2$ reprezintă suma pătratelor diferențelor verticale dintre valorile estimate (prezise) prin dreapta de regresie și valoarea medie a variabilei dependente. Această variație ne indică cât de bun este modelul de regresie comparativ cu simpla utilizare a mediei variabilei Y. SSR măsoară ponderea din totalul variației valorilor

Scanned with OKEN Scanner

observate ale variabilei Y care este datorată legăturii liniare dintre cele două variabile X și Y.

Sintetizând avem:

$$SST = SSE + SSR$$

Definirea mediilor sumelor pătratelor diferențelor verticale care exprimă modelul de regresie (MSR) și eroarea (MSE) se realizează prin raportarea sumelor prezentate (SSR și SSE) la numărul gradelor de libertate corespunzător.

În situația când presupunerile modelului de regresie se consideră îndeplinite, raportul mediilor sumelor pătratelor diferențelor (estimatorii variațelor) urmează o distribuție de tip Fisher cu 1 grad de libertate la numărător (numărul de variabile independente), respectiv $n-2$ grade de libertate la numitor:

$$F_{calculat} = \frac{MSR}{MSE} = \frac{\sum_{i=1}^{k}(\hat{y}_i - \bar{y})^2}{\frac{\sum_{i=1}^{k}(y_i - \hat{y}_i)^2}{n-2}} \quad (11.15)$$

În cazul regresiei simple liniare, testul ANOVA este utilizat pentru a verifica dacă modelul de regresie este semnificativ din punct de vedere statistic sau variabila independentă X are un efect semnificativ asupra variabilei dependente Y. Testul Fisher folosit este în acest caz un test omnibus, care verifică dacă coeficientul de regresie este egal cu zero în populația generală.

Tabelul 11.3. Tabelul ANOVA în cazul regresiei simple liniare

Sursa variației	**Suma pătratelor**	**Gradele de libertate**	**Estimatorii varianței populației**	**Raportul varianțelor (Statistica Fisher)**
Modelul de regresie	SSR	1	$MSR = SSR$	$F_{calculat} = \frac{MSR}{MSE}$
Erorile (val. reziduale)	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$	
Total	$\boldsymbol{SST}$	$\boldsymbol{n-1}$		

În cazul regresiei simple liniare, testul ANOVA verifică ipoteza nulă $H_0: \beta = 0$ (nu există nicio relație între variabilele X și Y) împotriva ipotezei alternative $H_1: \beta \neq 0$ (variabila X explică semnificativ variația variabilei Y)

Valoarea calculată a raportului varianțelor din tabelul 11.3, $F_{calculat}$, se compară cu valoarea din tabelul distribuției Fisher corespunzătoare gradelor de

Scanned with OKEN Scanner

libertate și unui nivel de semnificație standard (de obicei $\alpha = 0{,}05$) $F_{\alpha,1,n-2}$ (valoarea critică).

- dacă $F_{calculat} \geq F_{critic}$ $(p < 0{,}05)$, raportul variațelor se situează în zona de respingere, în extremitatea dreaptă a distribuției Fisher și în consecință vom respinge ipoteza nulă ($H_0: \beta = 0$) și vom concluziona că modelul liniar de regresie este bine adaptat datelor noastre;
- dacă $F_{calc} < F_{critic}$ $(p > 0{,}05)$ considerăm că nu avem suficiente dovezi pentru a respinge ipoteza nulă ($H_0: \beta = 0$), panta dreptei de regresie este egală cu zero și în consecință nu există nicio relație semnificativă între variabilele X și Y la nivel populațional.

Exemplu: folosim din nou programul Excel (meniul *„Data"*, modulul „Data Analysis", opțiunea *„Regression"*, pe datele prezentate anterior și obținem tabelul ANOVA pentru modelul nostru de regresie (variabila X - durata tratamentului cu corticosteroizi și variabila Y - nivelul glicemiei à jeun):

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	11524,493	11524,493	135,261	0,000
Residual	48	4089,687	85,202		
Total	**49**	**15614,180**			

În urma aplicării testului ANOVA pentru modelul de regresie simplă liniară considerat, s-a obținut o valoare a probabilității asociate (*Significance F*) mai mică decât nivelul de semnificație stabilit ($\alpha = 0{,}05$).

Deoarece $F_{calculat} = 135{,}261 > F_{critic} = F_{0.05,1,48} = 4{,}043$ $(p < 0{,}05)$, raportul variațelor se situează în zona de respingere, în extremitatea dreaptă a distribuției și în consecință avem suficiente dovezi în sprijinul respingerii ipotezei nule ($H_0: \beta = 0$) pentru un nivel de semnificație de 5%. Vom concluziona astfel că modelul liniar de regresie este semnificativ statistic și oferă o aproximare bună a datelor noastre (test unilateral dreapta). Cu alte cuvinte, există suficiente dovezi pentru a susține ideea că variația variabilei independente X - durata tratamentului cu corticosteroizi explică o parte semnificativă din variația variabilei dependente Y - nivelul glicemiei à jeun și, în consecință, este justificată folosirea modelului de regresie simplă liniară obținut în scop predictiv.

11.7. COEFICIENTUL DE DETERMINARE SIMPLĂ LINIARĂ

Este deosebit de utilă în practică evaluarea modelului de regresie simplă liniară prin calcularea raportului dintre variația explicată de modelul de regresie SSR și variația totală a modelului SST. Acest lucru se realizează prin

Scanned with OKEN Scanner

determinarea coeficientului de determinare (r^2), care evidențiază proporția (procentul) din variația variabilei dependente Y care este explicată de variația variabilei independente X:

$$r^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{k}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{k}(y_i - \bar{y})^2} \tag{11.16}$$

Coeficientul de determinare r^2 poate lua valori cuprinse între 0 (modelul de regresie nu explică deloc variația lui Y) și 1 (modelul de regresie explică întreaga variație a variabilei dependente Y, sau toate observațiile sunt situate pe dreapta de regresie). În practică de exemplu, atunci când r^2 ia valoarea 0,80 putem afirma faptul că 80% din variația valorilor variabilei dependente Y este explicată de variabila independentă X, iar restul de 20% reflectă influența altor factori, neluați în considerare în cadrul modelului.

Coeficientul de determinare r^2 calculat pe baza datelor din eșantion reprezintă un estimator punctual al coeficientului de determinare (ρ^2) din populația generală. În situația când numărul gradelor de libertate este redus (eșantioane de volum mic), r^2 prezintă o deplasare la dreapta (tinde să fie ușor mai ridicat ca valoare). Pentru a obține un estimator nedeplasat al lui ρ^2 în această situație se folosește varianta ajustată a coeficientului:

$$r^2_{adj} = 1 - \frac{\frac{\sum_{i=1}^{k}(\hat{y}_i - \bar{y})^2}{n-2}}{\frac{\sum_{i=1}^{k}(y_i - \bar{y})^2}{n-1}} \tag{11.17}$$

Exemplu: vom determina valoarea coeficientului de determinare pentru modelul de regresie simplă liniară prezentat anterior, utilizând programul Excel (meniul „*Data*", modulul „Data Analysis", opțiunea „*Regression*"):

Regression Statistics	
Multiple R	0,86
R Square	**0,74**
Adjusted R Square	0,73
Standard Error	9,23
Observations	50

Valoarea coeficientului de determinare $r^2 = 0{,}74$ evidențiază faptul că aproximativ 74% din variația variabilei dependente Y (nivelul glicemiei à jeun) poate fi explicată prin variația variabilei independente X (durata tratamentului cu corticosteroizi), iar restul de 26% din variație este determinat de alți factori

Scanned with OKEN Scanner

neincluși în model (factori genetici, dietă, alte tratamente, etc.) sau de posibile de erori de măsurare. Putem astfel concluziona că durata tratamentului cu corticosteroizi are un efect semnificativ și consistent asupra nivelului glicemiei à jeun. Modelul de regresie simplă liniară poate fi util în realizarea de predicții sau în monitorizarea pacienților care urmează tratamente pe bază de cortizon.

În plus, $r^2 = 0{,}74$ nu indică o relație cauzală între cele două variabile. Deși este posibil să existe un mecanism fiziologic cunoscut (tratamentele cu corticosteroizi pot conduce la creșterea nivelului glicemiei), modelul de regresie prezentat descrie doar asocierea statistică dintre variabile. Regresia evidențiază o relație matematică între variabile, nu un mecanism cauzal. Astfel, chiar dacă variația unei variabile este însoțită de o variație similară a celeilalte (corelație directă) acest lucru nu înseamnă că una o determină pe cealaltă. Principala problema care apare în acest context este reprezentată de faptul că, modelul de regresie simplă liniară nu controlează influența altor factori, cum ar fi: durata tratamentului, alte medicamente care pot interfera cu medicația de bază, dieta, sexul și vârsta pacienților, etc.). Pentru a putea vorbi de cauzalitate, este necesară prezența simultană a trei elemente principale: fundamentul logic al fenomenului medical, existența unei relații de asociere între variabile și controlul variabilelor care pot genera confuzie.

11.8. PRESUPUNERILE MODELULUI DE REGRESIE SIMPLĂ LINIARĂ

Presupunerile modelului de regresie simplă liniară reprezintă condițiile esențiale care trebuie îndeplinite anterior aplicării procedurii de inferență statistică privind estimarea modelului de regresie. Verificarea în prealabil a presupunerilor modelului de regresie liniară simplă este esențială pentru a asigura validitatea rezultatelor analizei și pentru a realiza o interpretare corespunzătoare a acestora.

Presupunerile modelului de regresie simplă liniară (figura 11.9):

1) pentru fiecare valoare x_i a variabilei independente X există mai multe valori corespunzătoare y_i ale variabilei dependente Y (*subpopulații* ale lui Y), care urmează *o distribuție normală*;
2) mediile acestor subpopulații distribuite normal ($\bar{y}_i$) se situează pe dreapta de regresie (*presupunerea de liniaritate*);
3) *valorile* y_i ale variabilei dependente Y *sunt independente*, în sensul că la momentul constituirii eșantionului valorile lui Y obținute pentru o anumită valoare x_i a lui X nu depind în niciun fel de valorile lui Y corespunzătoare altor valori ale lui X;
4) *deviațiile standard ale acestor distribuții normale sunt egale*, iar cea mai bună estimație a acestei deviații standard comune este eroarea standard

Scanned with OKEN Scanner

a estimării (o măsură a împrăștierii datelor observate în jurul dreptei de regresie, pentru o anumită valoare x_i a variabilei independente X):

$$s_{y.x} = \sqrt{\frac{SSE}{n-2}} = \sqrt{MSE} = \sqrt{\frac{\sum(y_i - \hat{y})^2}{n-2}} \tag{11.18}$$

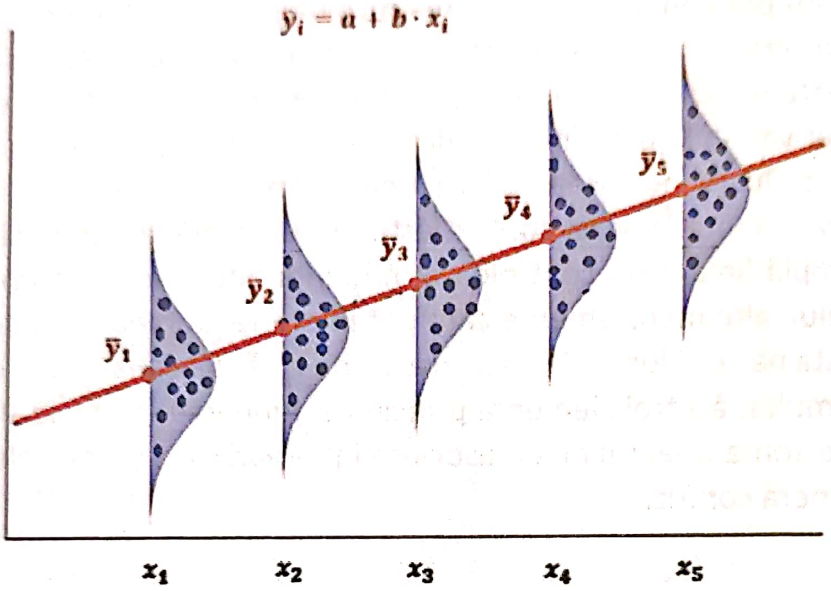


Figura 11.9. Reprezentarea grafică a presupunerilor modelului de regresie simplă liniară

5) modelul de regresie *nu trebuie să fie puternic influențat de prezența unor valori extreme* (outlieri), care pot modifica radical poziția dreptei de regresie. Pentru a estima influența valorilor outlierilor care pot distorsiona rezultatele analizei de regresie se poate calcula „distanța lui Cook" (D_i). Aceasta se obține eliminând valoarea observației i (potențialul outlier) din model și apoi refăcând analiza de regresie. Indicatorul exprimă cât de mult se modifică toate valorile din modelul de regresie atunci când observația i a fost eliminată. În practică, o valoare a distanței lui Cook peste 1 se consideră mare.

În analizele practice, valorile reziduale $\varepsilon_i = y_i - \hat{y}_i$ sunt utilizate în vederea testării *presupunerilor de liniaritate și de egalitate a variațelor* care stau la baza modelului de regresie. Astfel, valorile reziduale $(y_i - \hat{y}_i)$ sau valorile reziduale standardizate $z_{y_i - \hat{y}_i} = \frac{(y_i - \hat{y}_i) - o}{\sigma_{y_i - \hat{y}_i}}$, după caz, vor fi reprezentate pe axa verticală, iar valorile prezise $\hat{y}_i$ ale lui Y prin utilizarea ecuației de regresie pe axa orizontală. În situația în care punctele sunt împrăștiate în mod uniform pe

Scanned with OKEN Scanner

suprafața graficului, în jurul dreptei orizontale $y_i - \hat{y}_i = 0$, presupunerile se consideră a fi îndeplinite.

Atunci când avem o relație liniară între variabile, erorile (valorile reziduale) reprezintă doar fluctuații întâmplătoare în jurul dreptei de regresie a modelului. Astfel, o schimbare a valorii variabilei rezultat Y datorită modificării cu o unitate a lui X trebuie să fie constantă pentru toate valorile variabilei independente X. Se realizează distribuția erorilor $(y_i - \hat{y}_i)$ pentru toate valorile estimate de model. În situația când se sesizează o altă formă grafică neliniară, aceasta poate fi un semnal al neliniarității datelor; de asemenea, dacă datele sunt dispuse în formă de pâlnie, avem un semn al unei varianțe crescătoare.

Varianța erorilor ε_i este constantă pentru toate nivelurile variabilei independente X; astfel, se presupune că fiecare valoare reziduală $\varepsilon_i = y_i - \hat{y}_i$ este generată dintr-o distribuție normală de medie zero și o deviație standard σ, care este aceeași pentru fiecare valoare reziduală).

Exemplu: vom verifica presupunerile modelului de regresie simplă liniară prezentat anterior. Construim graficul de tip „nor de puncte" (scatter-plot) al valorilor reziduale standardizate și observăm o distribuție întâmplătoare a punctelor pentru toată plaja de variație a variabilei dependente estimate, care este reprezentată pe axa absciselor. Punctele sunt complet împrăștiate, fără a se putea evidenția niciun fel de model care să pună la îndoială presupunerea de liniaritate (figura 11.10).

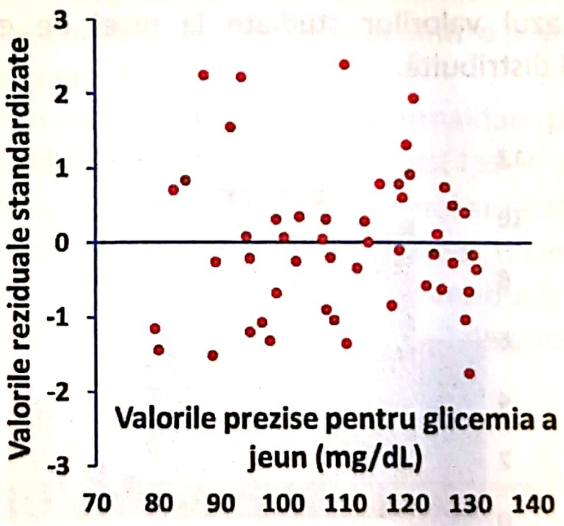


Figura 11.10. Distribuția valorilor reziduale standardizate

De asemenea, observăm faptul că, variația valorilor reziduale este constantă, indiferent dacă valorile variabilei dependente estimate prin regresie sunt mari sau mici și astfel putem concluziona că variabilitatea în jurul ecuației de regresie

Scanned with OKEN Scanner

este aceeași pentru toate valorile variabilei independente (condiția de homoscedasticitate este îndeplinită – figura 11.10).

Pentru a putea conchide faptul că analiza statistică inferențială privind modelul de regresie liniară este validă, trebuie evaluată distribuția valorilor reziduale, care în condiții ideale ar trebui să urmeze o repartiție normală, având media zero. Prin urmare, vom calcula valorile reziduale standardizate z_{ε_i} pentru fiecare valoare estimată a variabilei dependente, prin raportarea valorilor reziduale ε_i la deviația standard comună a acestora σ_{ε_i} (tabelul 11.4).

Tabelul 11.4. Determinarea valorilor reziduale standardizate ale variabilei Y

Pacient	y_i	$\hat{y}_i$	$\varepsilon_i = y_i - \hat{y}_i$	z_{ε_i}
1	118,000	118,883	-0,883	-0,097
2	75,000	88,773	-13,773	-1,508
3	133,000	129,330	3,670	0,402
4	116,000	113,353	2,647	0,290
...	...	...	...	...

Ulterior, vom reprezenta grafic distribuția de frecvențe a valorilor reziduale standardizate z_{ε_i} prin construirea histogramei lor (figura 11.11). După cum remarcăm din analiza figurii 11.11 distribuția valorilor reziduale standardizate în cazul valorilor studiate la nivel de eșantion pare să fie aproximativ normal distribuită.

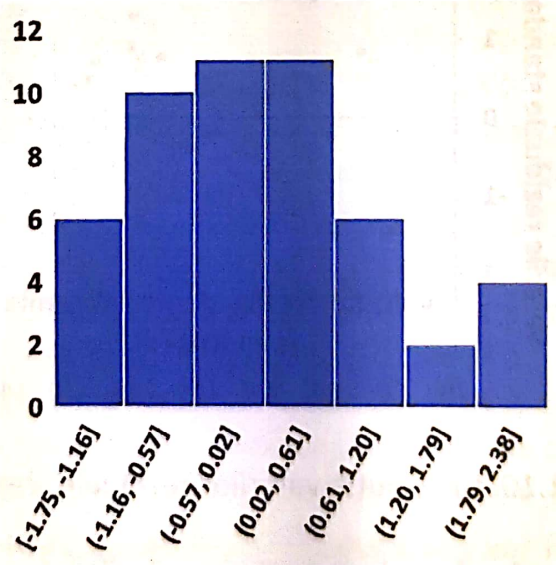


Figura 11.11. Histograma valorilor reziduale standardizate

Scanned with OKEN Scanner

Pentru a stabili din punct de vedere vizual dacă setul de date provine dintr-o distribuție statistică gausiană, vom construi diagrama Q-Q Plot a valorilor reziduale standardizate (figura 11.12).

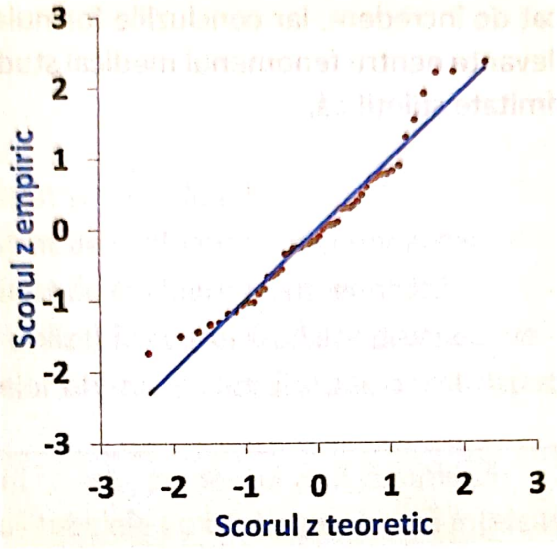


Figura 11.12. Diagrama Q-Q Plot a valorilor reziduale standardizate

Analizând datele prezentate în figura 11.12, observăm că punctele nu deviază foarte mult de la linia de referință înclinată la 45° (cu câteva mici excepții) și în consecință putem concluziona faptul că valorile reziduale standardizate determinate la nivelul eșantionului studiat provin dintr-o populație care urmează o distribuție aproximativ normală.

În final, vom folosi și un test statistic de normalitate pentru distribuția de frecvențe a valorilor reziduale standardizate z_{ε_i}. Acest test evaluează ipoteza nulă conform căreia populația din care provine eșantionul studiat ar putea fi normal distribuită. Generăm cu ajutorul programului SPSS testele de verificare a normalității distribuției de frecvențe a valorilor reziduale standardizate din populația generală, din care provine eșantionul de 50 de pacienți:

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Glicemie_mg_dL	.097	50	.200*	.962	50	.103

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Având în vedere faptul că, în situația ambelor teste de verificare a normalității, valoarea probabilității $p \geq 0{,}05$, concluzionăm că valorile reziduale

Scanned with OKEN Scanner

standardizate ale populației generale, din care a fost extras în mod aleatoriu eșantionul studiat, ar putea fi normal distribuite.

Întrucât presupunerile de bază ale modelului de regresie simplă liniară sunt îndeplinite, relația dintre cele două variabile studiate poate fi interpretată cu un nivel mai ridicat de încredere, iar concluziile formulate în baza modelului pot fi considerate relevante pentru fenomenul medical studiat, căpătând în acest mod mai multă legitimitate științifică.

Scanned with OKEN Scanner

Capitolul. 12

12. TESTAREA ASOCIERII DINTRE VARIABILELE CATEGORIALE

SUBIECTE:

12.1. Testul CHI-pătrat și formele lui
12.2. Testul CHI-pătrat de evaluare a independenței
12.3. Testul CHI-pătrat de evaluare a omogenității
12.4. Riscul relativ utilizat în cadrul studiilor prospective
12.5. Raportul cotelor utilizat în cadrul studiilor retrospective

La finalizarea acestui capitol, studentul va fi capabil să:
O12-1: să definească tabelele de contingență și să înțeleagă importanța lor în analiza statistică
O12-2: să poată utiliza testul CHI-pătrat pentru a testa independența sau omogenitatea
O12-3: să poată aplica testul exact al lui Fisher în cazul tabelelor 2 x 2
O12-4: să înțeleagă modul de calcul precum și interpretarea riscului relativ
O12-5: să înțeleagă modul de calcul precum și interpretarea raportului cotelor

12.1. TESTUL CHI-PĂTRAT ȘI FORMELE LUI

Distribuția CHI-pătrat reprezintă cea mai utilizată tehnică de analiză statistică în cazul variabilelor categoriale sau calitative (mediul social, starea civilă, genul, grupa de sânge, etc.) care se bazează pe studierea frecvențelor asociate cu fiecare categorie a variabilelor studiate. De exemplu, analizând un eșantion de pacienți pentru care cunoaștem distribuția frecvențelor tipurilor de diagnostic cât și împărțirea pe grupe sanguine, am dori să aflăm dacă, pentru întreaga populație din care provine acest lot, există o relație (asociere sau legătură) între grupa de sânge a pacienților și diagnostic.

În acest context, tabelele de contingență reprezintă un instrument esențial, folosit în principal pentru a organiza, descrie și cerceta relațiile care ar putea exista între două sau mai multe variabile categoriale (calitative). Metoda de analiză statistică este deosebit de utilă în domeniul medical, unde oferă posibilitatea evidențierii unei potențiale asocieri între un factor de risc și o anumită afecțiune sau posibilitatea investigării performanței unui test de

Scanned with OKEN Scanner

diagnostic, contribuind în acest fel la luarea unor decizii clinice informate, la evaluarea riscurilor și la validarea ipotezelor în cercetarea științifică medicală.

Un *tabel de contingență*, cunoscut și sub numele de tabel cu frecvențe încrucișate sau tabel cu dublă intrare, reprezintă o matrice a distribuției de frecvențe bidimensionale, în care sunt prezentate frecvențele absolute (sau relative) ale tuturor combinațiilor posibile dintre două variabile statistice categoriale. Structurarea datelor statistice în cazul variabilelor categoriale sub forma tabelelor de contingență stă la baza unor teste de verificare a ipotezelor statistice importante, cum ar fi de exemplu testul χ^2, introdus de Karl Pearson în anul 1900, inițial pentru a evalua dacă o serie de observații urmează o anumită distribuție teoretică sau testul exact al lui Fisher, propus de către Sir Ronald Fisher în anul 1934, pentru a compara proporțiile între două grupuri de volum redus, în situația în care testul χ^2 nu returnează rezultate valide.

Tabelul 12.1. Macheta unui tabel de contingență, cu r linii și c coloane

Variabila Y / Variabila X	y_1	y_2	...	y_j	...	y_c	Total
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1c}	$n_{1.} = \sum_{j=1}^{c} n_{1j}$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2c}	$n_{2.} = \sum_{j=1}^{c} n_{2j}$
...	...	...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ic}	$n_{i.} = \sum_{j=1}^{c} n_{ij}$
...	...	...	...	...	...	...	...
x_r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rc}	$n_{r.} = \sum_{j=1}^{c} n_{rj}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.j}$	...	$n_{.c}$	$n = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}$

Studiul pornește prin selectarea în mod aleatoriu a unui eșantion dintr-o populație de interes, fiecare caz înregistrat fiind clasificat simultan în funcție de două variabile statistice (de exemplu, variabila X poate fi factorul de risc – statusul fumător/nefumător al fiecărui subiect, iar variabila Y prezența sau

Scanned with OKEN Scanner

absența maladiei – cancerul pulmonar). Ulterior, datele statistice colectate experimental sunt grupate în tabele de contingență, care cuprind toate combinațiile dintre variantele (categoriile) pe care le au cele două variabile calitative studiate X și Y. Clasificarea simultană a unităților statistice studiate în funcție de două criterii poate fi prezentată sub formă tabelară (figura 12.1), cele r linii reprezentând nivelurile (categoriile) primei variabile, iar cele c coloane nivelurile celei de a doua.

În cadrul analizei statistice inferențiale a asocierii dintre aceste două variabile categoriale vom testa dacă în cadrul populației generale de volum N, din care provine eșantionul studiat de volum n, cele două variabile statistice pot fi considerate independente. Două variabile statistice se consideră independente în situația în care între acestea nu există nicio relație consistentă sau predictibilă între acestea (distribuția frecvențelor primei variabile nu depinde în niciun fel de categoriile celei de a doua).

Există două tipuri principale de frecvențe de interes în cadrul acestei cercetări statistice: frecvențele observate și frecvențele așteptate. Toate valorile n_{ij} prezentate în tabelul de contingență inițial reprezintă numărul real de cazuri (persoane sau obiecte) din fiecare combinație de categorii ale variabilelor analizate sau *frecvențele observate*. De exemplu, frecvența n_{ij} reprezintă numărul de cazuri din eșantion care poartă simultan atât nivelul sau categoria i a variabilei X cât și nivelul j al variabilei Y. A doua categorie, respectiv *frecvențele așteptate*, evidențiază numărul de cazuri din eșantion pe care ne așteptăm să le avem în condițiile în care ipoteza nulă H_0 a testului statistic care analizează cele două variabile ar fi adevărată. În acest fel, frecvențele așteptate definesc o distribuție ideală, ipotetică, care este în deplină concordanță cu ipoteza nulă.

Testele statistice bazate pe distribuția CHI-pătrat (χ^2) se obțin pe baza următoarei formule generale de calcul:

$$\chi^2_{calculat} = \sum_{i=1}^{k} \left[\frac{(Frecventa\ observata_i - Frecventa\ asteptata_i)^2}{Frecventa\ asteptata_i} \right] \quad (12.1)$$

În situația în care ipoteza nulă H_0 este adevărată, valoarea calculată a testului urmează o distribuție aproximativă χ^2, cu $(r-1)\cdot(c-1)$ grade de libertate. Nivelul $\chi^2_{calculat}$, determinat pe baza eșantionului studiat, reprezintă o măsură a concordanței generale existente între perechile de valori ale frecvențelor observate și ale celor așteptate. Prin urmare, în cazul când frecvențele observate și cele așteptate sunt relativ asemănătoare valoarea $\chi^2_{calculat}$ va fi redusă, iar atunci când diferențele sunt importante valoarea va fi ridicată. Problema care se pune în această situație este cât de mare ar trebui să fie nivelul lui $\chi^2_{calculat}$ pentru a putea în final respinge ipoteza nulă. Pentru a nu pierde din informație, oferindu-le în acest mod ponderi egale, diferențele dintre

Scanned with OKEN Scanner

frecvențele observate și cele așteptate, de la numărătorul relației 12.1, au fost ridicate la pătrat înainte de însumarea lor. Ulterior, prin împărțirea la nivelurile frecvențelor așteptate corespunzătoare, vom obține un indicator statistic exprimat în unitățile originale de măsură.

Această valoare a testului calculat la nivel de eșantion va fi în final comparată cu valoarea teoretică din tabelul distribuției CHI-pătrat, cu $(r-1)\cdot(c-1)$ grade de libertate. În final, vom respinge ipoteza nulă H_0 în situația când valoarea calculată a testului nostru statistic este superioară valorii teoretice din tabel ($\chi^2_{calculat} > \chi^2_{tabel}$), pentru un anumit nivel de semnificație α ales în cadrul studiului (de regulă $\alpha = 0{,}05$ sau $\alpha = 0{,}01$).

În cadrul analizelor practice, testul CHI-pătrat poate avea două variante principale:

- *test de independență*: datele colectate sunt privite ca aparținând unui singur eșantion, extras din aceeași populație generală, pentru fiecare individ înregistrându-se nivelurile ambelor variabile; în acest caz, scopul final este reprezentat de evaluarea relației care ar putea exista la nivel populațional între variabilele de interes X și Y; de exemplu, dacă variabilele hipertensiune (cu categoriile DA și NU) și diabet (cu categoriile DA și NU) într-o anumită populație sunt independente (ipoteza nulă H_0 este adevărată), ne vom aștepta să găsim aceeași proporție de persoane „cu hipertensiune" și „fără hipertensiune" în ambele grupe de persoane „cu diabet" și „fără diabet";
- *test de omogenitate*: datele colectate sunt privite ca reprezentând două sau mai multe eșantioane separate, care aparțin unor populații diferite, scopul testului CHI-pătrat fiind în acest caz evaluarea existenței unor potențiale diferențe semnificative statistic între populațiile comparate. Întrebarea la care trebuie să răspundem în acest caz este: sunt eșantioanele extrase din cadrul unor populații care sunt omogene din punct de vedere al criteriului de clasificare (variabilei de interes)? De exemplu, se dorește studierea frecvenței de apariție a migrenelor în cazul a 2 eșantioane de pacienți, extrase în mod aleatoriu din cadrul a două populații diferite: 75 diagnosticați cu narcolepsie și 75 de persoane sănătoase. În urma derulării studiului se dorește să se cunoască dacă populațiile de narcoleptici și de persoane sănătoase, din care provin aceste 2 eșantioane, sunt omogene din punct de vedere a frecvenței migrenelor;

Scanned with OKEN Scanner

12.2. TESTUL CHI-PĂTRAT DE EVALUARE A INDEPENDENȚEI

Varianta cel mai des întâlnită în practică a distribuției CHI-pătrat este reprezentată de testarea ipotezei nule a independenței dintre două variabile statistice categoriale care caracterizează un set de entități studiate (persoane, obiecte, fenomene, etc.), provenind din aceeași populație de referință. De exemplu, în situația în care calitatea de fumător/nefumător a unei persoane și prezența/absența hipertensiunii arteriale pentru indivizii aparținând aceleiași populații de interes sunt variabile independente ne așteptăm la procente relativ apropiate ale persoanelor fumătoare și nefumătoare în cele două grupuri, respectiv cu și fără hipertensiune arterială. Cu alte cuvinte, indiferent dacă o persoană fumează sau nu, aceasta poate prezenta sau nu hipertensiune arterială, între cele două variabile categoriale neexistând nicio relație consistentă sau predictibilă care să poată fi evidențiată.

Ipotezele testului CHI-pătrat (χ^2) de evaluare a independenței dintre două variabile categoriale vor fi:

- *Ipoteza nulă* H_0: între variabilele categoriale studiate X și Y nu există nicio asociere la nivelul populației generale din care eșantionul a fost extras în mod aleatoriu (variabilele sunt independente);
- *Ipoteza alternativă* H_1: există o asociere la nivel populațional între variabilele X și Y (variabilele nu sunt independente).

Testul χ^2 nu oferă nicio informație asupra relației de cauzalitate dintre două variabile categoriale, acesta reprezentând doar o măsură a intensității asocierii între acestea.

Formula generală de calcul a testului χ^2 este cea prezentată anterior în relația 12.1., frecvențele observate fiind valorile absolute determinate la nivelul eșantionului studiat sau numărul de cazuri care respectă anumite condiții, iar frecvențele așteptate reprezentând valorile determinate teoretic pentru situația ipotetică în care nu ar exista nicio legătură între cele 2 variabile (ipoteza H_0 este adevărată).

Cel mai simplu caz este reprezentat de testarea independenței dintre 2 variabile categoriale binare (dihotomice), care pot lua fiecare doar 2 valori: Da/Nu, Adevărat/Fals, 1/0. Datele pot fi sistematizate sub forma tabelului 12.2.

Frecvențele așteptate pot fi determinate printr-o abordare probabilistică. Astfel în prima celulă a tabelului vom avea probabilitatea ca ambele variabile să înregistreze nivelul „DA":

Probabilitatea ca variabila 1 să înregistreze nivelul „DA" (evenimentul 1) este:

$$P_1 = P(Variabila\ 1 = "DA") = \frac{(a+b)}{n} \tag{12.2}$$

Scanned with OKEN Scanner

Tabelul 12.2. Frecvențele observate în cazul a două variabile dihotomice

Frecvențele observate		Variabila 2		Total
		DA	NU	
Variabila 1	DA	a	b	$a+b$
	NU	c	d	$c+d$
Total		$a+c$	$b+d$	$a+b+c+d=n$

Probabilitatea ca variabila 2 să înregistreze nivelul „DA" (evenimentul 2) este:

$$P_2 = P(Variabila\ 2 = "DA") = \frac{(a+c)}{n} \tag{12.3}$$

Deoarece lucrăm în condiția unei ipoteze nule H_0 adevărate, rezultă că evenimentele sunt considerate independente, astfel încât probabilitatea realizării simultane a lor va fi egală cu produsul probabilităților producerii fiecăruia în parte:

$$P = P_1 \cdot P_2 = \frac{(a+b)\cdot(a+c)}{n\cdot n} \tag{12.4}$$

Probabilitatea din formula 12.4 poate fi privită ca reprezentând de fapt frecvența relativă a acestui eveniment, care poate lua valori între 0 și 1. Pentru a obține valoarea frecvenței absolute corespunzătoare acesteia, sau frecvența absolută așteptată, vom înmulți această valoare cu numărul total de cazuri (n):

$$Frecventa\ asteptata = \frac{(a+b)\cdot(a+c)}{n\cdot n}\cdot n = \frac{(a+b)\cdot(a+c)}{n} \tag{12.5}$$

Prin urmare, frecvențele absolute așteptate, corespunzătoare celor observate, se obțin astfel:

$$Frecventa\ asteptata = \frac{Totalul\ liniei \cdot Totalul\ coloanei}{Totalul\ general} \tag{12.6}$$

Ulterior determinării tuturor valorilor frecvențelor așteptate, conform relației 12.6, datele se vor sistematiza sub forma tabelului 12.3:

Totalurile marginale, în cazul ambelor variabile, respectiv $a+b$ și $c+d$ pentru variabila 1 precum și $a+c$ și $b+d$ pentru variabila 2, sunt identice în ambele cazuri, atât pentru frecvențele observate și pentru cele așteptate (tabelele 12.2 și 12.3).

Scanned with OKEN Scanner

Tabelul 12.3. Frecvențele așteptate în cazul a două variabile dihotomice

Frecvențele așteptate		Variabila 2		Total
		DA	NU	
Variabila 1	DA	$\frac{(a+b)\cdot(a+c)}{n}$	$\frac{(a+b)\cdot(b+d)}{n}$	$a+b$
	NU	$\frac{(c+d)\cdot(a+c)}{n}$	$\frac{(c+d)\cdot(b+d)}{n}$	$c+d$
Total		$a+c$	$b+d$	$a+b+c+d=n$

Frecvențele așteptate vor fi comparate cu cele observate la nivel de eșantion și, în situația unor discrepanțe suficient de mici, ipoteza nulă a absenței asocierii dintre variabile este considerată rezonabilă. Dacă însă diferențele sunt suficient de mari, ipoteza nulă va fi respinsă în favoarea ipotezei alternative. Valoarea calculată a testului definit în această manieră urmează o aproximativ o distribuție χ^2 cu $(randuri-1)\cdot(coloane-1)$ grade de libertate, în situația când ipoteza nulă este adevărată. Atunci când $\chi^2_{calculat} > \chi^2_{tabel,\ df}$ respingem ipoteza nulă a absenței asocierii dintre cele două variabile pentru un nivel α de semnificație și df grade de libertate și concluzionăm faptul că există o relație între cele două variabile categoriale la nivelul populației generale, din care provine eșantionul studiat.

În cazul tabelelor de contingență $(2\ x\ 2)$ avem în general 4 grade de libertate, fiecare celulă conținând un număr. Atunci când totalurile marginale ale tabelului sunt însă precizate, rămâne doar *un singur grad de libertate* (e nevoie doar de o singură valoare din cele 4, restul putând fi calculate în funcție de totalurile liniilor, coloanelor și respectiva valoare).

Pentru a obține rezultate valide în urma aplicării testului χ^2 este necesară verificarea îndeplinirii unor condiții preliminare:

- variabilele studiate trebuie să fie de tip categorial;
- fiecare observație trebuie să fie independentă de toate celelalte;
- volumul eșantionului trebuie să fie suficient de mare, astfel încât valorile frecvențelor așteptate să fie cel puțin egale cu 5;
- fiecare unitate statistică colectată la nivel de eșantion trebuie să aparțină doar unei singure categorii.

Exemplu: cercetam posibila asociere între factorul de risc - statusul (fumător / nefumător) și maladie - prezenta cancerului pulmonar, pentru un eșantion de 100 de persoane. Datele colectate la nivel de eșantion sunt prezentate în tabelul 12.4:

Scanned with OKEN Scanner

Tabelul 12.4. Tabelul cu datele inițiale ale celor 100 de pacienți din eșantion

Pacient	Factor de risc (status fumător/nefumător)	Maladie (cancer pulmonar)
1	NU	NU
2	DA	DA
3	DA	NU
...	...	...
99	DA	NU
100	DA	DA

Inițial realizăm o grupare a datelor după cele 2 variabile de interes și o reprezentare grafică a acestora. Realizarea tabelului de contingență pentru datele din tabelul 12.4 se poate realiza utilizând modului „Pivot Table" inclus în programul Excel sau opțiunea „Crosstabs" din sub-meniul „Descriptive Statistics" al programului IBM SPSS. Datele prelucrate privind frecvențele observate sunt prezentate în tabelul 12.5:

Tabelul 12.5. Frecvențele observate ale celor două variabile

Frecvențele observate	**Maladia**		
Factor de risc	**DA**	**NU**	**TOTAL**
DA	38,0	15,0	**53,0**
NU	14,0	33,0	**47,0**
TOTAL	**52,0**	**48,0**	**100,0**

Informațiile prezentate centralizat, sub forma tabelului de contingență 12.5 se pot reprezenta în manieră grafică utilizând graficul tip „stivă de bare", care oferă o vizualizare foarte clară a distribuției categoriilor într-un mod intuitiv și ușor de interpretat. Fiecare bară orizontală, reprezentând categoriile primei variabile este împărțită în segmente care corespund categoriilor celei de a doua variabile. Printre avantajele acestui tip de grafic se numără evidențierea clară a structurii interne a fiecărui grup, aspect care permite realizarea unei comparații facile a ponderilor deținute de fiecare categorie.

Scanned with OKEN Scanner

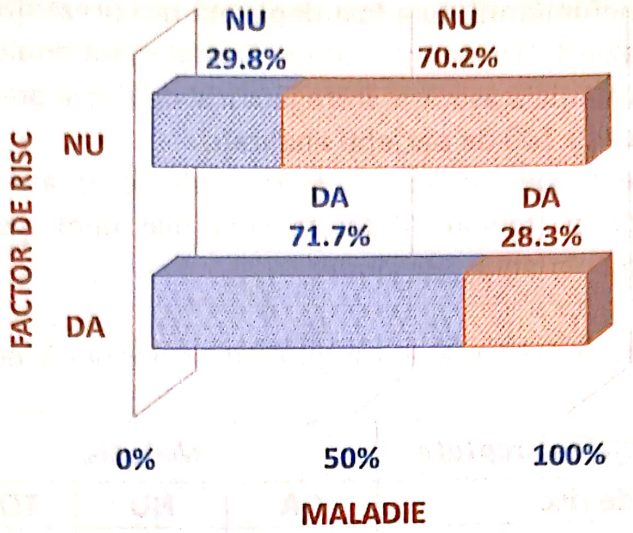


Figura. 12.1. Reprezentarea frecvențelor observate prin utilizarea graficului tip „stivă de bare"

În analizele practice, se mai utilizează și graficul prin „coloane grupate", care evidențiază separat, valorile absolute ale categoriilor celor două variabile (figura 12.2).

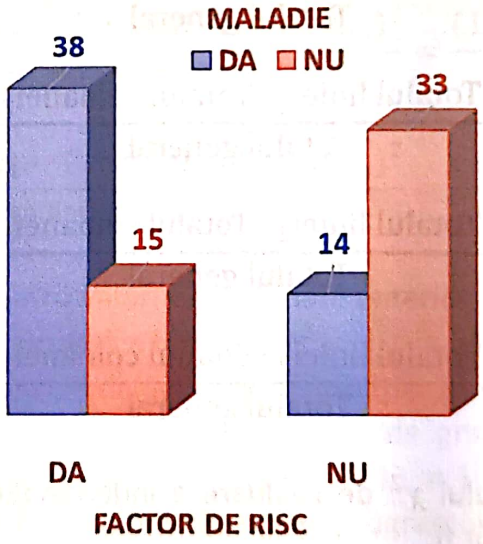


Figura. 12.2. Reprezentarea frecvențelor observate prin utilizarea graficului prin „coloane grupate"

Atât din tabelul 12.5 cât și din cele două reprezentări grafice aferente (figurile 12.1 și 12.2) se remarcă imediat faptul că la pacienții fumători (supuși acțiunii factorului de risc) maladia (cancerul pulmonar) este prezentă în proporție de 71,7% (38 din 53 de cazuri), iar la majoritatea pacienților nesupuși acțiunii

Scanned with OKEN Scanner

factorului de risc (nefumători) nu a fost depistată nici prezența maladiei (70,2% sau 33 din 47 de cazuri). Datele colectate la nivelul eșantionului de lucru par să sugereze astfel o potențială asociere între factorul de risc și prezența maladiei în cazul eșantionului celor 100 de pacienți analizați.

În continuare vom determina frecvențele așteptate corespunzătoare frecvențelor observate (tabelul 12.6), în condițiile unei ipoteze nule H_0 a independenței dintre variabile adevărate:

Tabelul 12.6. Frecvențele așteptate ale celor două variabile

Frecvențele așteptate	Maladia		
Factor de risc	DA	NU	TOTAL
DA	27,6	25,4	53,0
NU	24,4	22,6	47,0
TOTAL	52,0	48,0	100,0

Frecvențele așteptate centralizate în tabelul 12.6 s-au calculat prin utilizarea relației 12.6, după cum urmează:

$$\text{Frecv. asteptata}_1 = \frac{\text{Totalul liniei}_1 \cdot \text{Totalul coloanei}_1}{\text{Totalul general}} = \frac{53{,}0 \cdot 52{,}0}{100} = 27{,}6$$

$$\text{Frecv. asteptata}_2 = \frac{\text{Totalul liniei}_1 \cdot \text{Totalul coloanei}_2}{\text{Totalul general}} = \frac{53{,}0 \cdot 48{,}0}{100} = 25{,}4$$

$$\text{Frecv. asteptata}_3 = \frac{\text{Totalul liniei}_2 \cdot \text{Totalul coloanei}_1}{\text{Totalul general}} = \frac{47{,}0 \cdot 52{,}0}{100} = 24{,}4$$

$$\text{Frecv. asteptata}_4 = \frac{\text{Totalul liniei}_2 \cdot \text{Totalul coloanei}_2}{\text{Totalul general}} = \frac{47{,}0 \cdot 48{,}0}{100} = 22{,}6$$

Ipotezele testului χ^2 de evaluare a independenței dintre aceste două variabile categoriale vor fi:

- *Ipoteza nulă* H_0: între variabilele studiate X (factorul de risc) și Y (maladia) nu există nicio asociere la nivelul populației generale din care eșantionul a fost extras în mod aleatoriu (variabilele sunt independente);
- *Ipoteza alternativă* H_1: există o asociere la nivel populațional între variabilele X și Y (variabilele nu sunt independente).

Scanned with OKEN Scanner

În final, vom realiza o comparație a frecvențelor observate cu cele așteptate, prin determinarea valorii statisticii $\chi^2_{calculat}$. Rezultatele obținute prin utilizarea programului Excel sunt prezentate în tabelul 12.7:

Tabelul 12.7. Determinarea practică a valorii testului χ^2

Test χ^2	Maladia		
Factor de risc	**DA**	**NU**	**TOTAL**
DA	3,96	4,28	8,24
NU	4,46	4,83	9,29
TOTAL	8,41	9,12	**17,53**

Cele patru valori din tabelul 12.7 au fost obținute în baza relației 12.1, astfel:

$$\frac{(\text{Frecventa observata}_1 - \text{Frecventa asteptata}_1)^2}{\text{Frecventa asteptata}_1} = \frac{(38{,}0 - 27{,}6)^2}{27{,}6} = 3{,}96$$

$$\frac{(\text{Frecventa observata}_2 - \text{Frecventa asteptata}_2)^2}{\text{Frecventa asteptata}_2} = \frac{(15{,}0 - 25{,}4)^2}{25{,}4} = 4{,}28$$

$$\frac{(\text{Frecventa observata}_3 - \text{Frecventa asteptata}_3)^2}{\text{Frecventa asteptata}_3} = \frac{(14{,}0 - 24{,}4)^2}{24{,}4} = 4{,}46$$

$$\frac{(\text{Frecventa observata}_4 - \text{Frecventa asteptata}_4)^2}{\text{Frecventa asteptata}_4} = \frac{(33{,}0 - 22{,}6)^2}{22{,}6} = 4{,}83$$

Însumând cele patru valori determinate anterior: $3{,}96 + 4{,}28 + 4{,}46 + 4{,}83$ obținem valoarea calculată a testului: $\chi^2_{calculat} = 17{,}53$.

În continuare, vom citi din tabelele distribuției χ^2 valoarea critică, pentru un nivel de semnificație $\alpha = 0{,}05$ și un număr de grade de libertate: $df = (randuri - 1) \cdot (coloane - 1) = (2 - 1) \cdot (2 - 1) = 1$. Astfel $\chi^2_{tabel,1} = 3{,}841$ (valoarea poate fi obținută direct în programul Excel, prin utilizarea formulei: $= CHISQ.INV.RT(0.05,1)$.

Întrucât $\chi^2_{calculat} = 17{,}530 > \chi^2_{tabel,1} = 3{,}841$, concluzionăm faptul că avem suficiente dovezi furnizate de eșantionul studiat, pentru a putea respinge ipoteza nulă a independenței dintre variabile. Considerăm astfel că, la nivelul populației generale din care a fost extras în mod aleatoriu eșantionul, ar putea exista o asociere semnificativă statistic între factorul de risc (fumatul) și maladie (cancerul pulmonar).

Scanned with OKEN Scanner

Valoarea probabilității asociate testului χ^2, care evidențiază probabilitatea de a obține o valoare cel puțin la fel de „extremă" cum este cea determinată la nivel de eșantion de 17,530, atunci când ipoteza nulă este adevărată, este $p = 0{,}00003$ (se determină în Excel cu formula: $= CHISQ.TEST(frecv.observate, frecv.așteptate)$. Prin urmare, probabilitatea de a respinge ipoteza nulă H_0 atunci când ea este adevărată, este extrem de redusă în cazul datelor din cadrul eșantionului studiat.

În situația unui volum redus al eșantionului de lucru, respectiv atunci când 20% dintre celule au frecvențele așteptate sub valoarea 5, în cadrul tabelelor de contingență cu mai mult de 1 grad de libertate, eroarea introdusă în calcul crește foarte mult. O soluție la această problemă ar putea fi reprezentată de *combinarea rândurilor sau coloanelor alăturate*, care ar conduce la obținerea unor valori așteptate ale frecvențelor mai mari. În cazul tabelelor de contingență 2 x 2, Cochran sugerează faptul că testul χ^2 nu ar trebui utilizat dacă $n < 20$ sau dacă $20 < n < 40$ și orice frecvență așteptată este mai mică decât 5. Atunci când $n \geq 40$ poate fi tolerată o singură frecvență așteptată (din cele 4) egală cu 1.

O altă soluție pentru tabelele de contingență 2 x 2 este reprezentată de *testul exact Fisher*, care permite calcularea probabilității exacte de a obține rezultatele observate sau rezultate „mai" extreme. Ipotezele testului exact Fisher sunt: cele 2 eșantioane sunt extrase în mod aleatoriu și sunt independente, iar fiecare observație poate fi încadrată doar într-una dintre cele două categorii exclusive ale fiecărei variabile.

Testul exact Fisher calculează numărul tuturor tabelelor de contingență posibile cu aceleași totaluri marginale (pe linii și coloane) ca tabelul inițial al frecvențelor observate. Ulterior se calculează probabilitatea p, determinând ponderea din numărul total al tabelelor deținută de tabelele mai neobișnuite comparativ cu tabelul inițial.

Frecvențele observate într-un tabel de contingență 2 x 2 reprezintă valori discrete și astfel dau naștere unei statistici calculate discrete, care este aproximată prin distribuția continuă χ^2. În anul 1934, statisticianul englez Frank Yates, unul dintre discipolii lui Sir Ronald Aylmer Fisher, a propus o procedură de corectare a acestei situații, care conduce la o aproximare mai bună a valorii statisticii calculate la nivel de eșantion, în cazul particular 2 x 2:

$$\chi^2_{calculat} = \sum_{i=1}^{k} \left[\frac{(|Frecventa\ observata_i - Frecventa\ asteptata_i| - 1/2)^2}{Frecventa\ asteptata_i} \right] \quad (12.7)$$

În situația utilizării unor eșantioane de volum mare, diferențele între valoarea testului χ^2 standard și cea obținută prin aplicarea corecției lui Yates devin neglijabile.

Scanned with OKEN Scanner

12.3. TESTUL CHI-PĂTRAT DE EVALUARE A OMOGENITĂȚII

Testul χ^2 de evaluare a omogenității este aplicat unei singure variabile categoriale din cadrul a două sau mai multe populații, pentru a determina dacă frecvențele observate sunt distribuite identic între acestea (de ex. se dorește să se afle dacă studenții de sex masculin și feminin au fost acceptați în procente diferite la specialitățile existente în cadrul unei facultăți).

Presupunând că eșantioanele provin din r populații, iar variabila categorială are c niveluri, în conformitate cu ipoteza nulă H_0, pentru fiecare nivel specificat al variabilei categoriale, fiecare populație are aceeași proporție a observațiilor. În acest context, pentru a decide dacă frecvențele observate la nivelul fiecărui eșantion diferă semnificativ de frecvențele așteptate conform ipotezei nule, vom aplica testul χ^2 de verificare a omogenității.

Testul χ^2 de verificare a omogenității este identic din punct de vedere a procedurii matematice de lucru cu testul χ^2 de evaluare a independenței.

Formula de calcul generală este:

$$\chi^2_{calculat} = \sum_{i=1}^{k} \left[\frac{(Frecventa\ observata_i - Frecventa\ asteptata_i)^2}{Frecventa\ asteptata_i} \right] \tag{12.8}$$

Frecvențele așteptate se obțin similar cu cazul anterior:

$$Frecventa\ asteptata = \frac{Totalul\ liniei * Totalul\ coloanei}{Totalul\ general} \tag{12.9}$$

Valoarea calculată a testului definit în această manieră urmează o aproximativ o distribuție χ^2 cu $(randuri - 1) \cdot (coloane - 1)$ grade de libertate, în situația când ipoteza nulă este adevărată.

Ipotezele testului CHI-pătrat (χ^2) de evaluare a omogenității sunt:

- *Ipoteza nulă* H_0: populațiile sunt considerate omogene (distribuite identic) din punctul de vedere al variabilei categoriale analizate;
- *Ipoteza alternativă* H_1: populațiile nu sunt considerate omogene din punctul de vedere al variabilei categoriale analizate.

Atunci când $\chi^2_{calculat} > \chi^2_{tabel,\ df}$ respingem ipoteza nulă a omogenității pentru un nivel α de semnificație și df grade de libertate și concluzionăm faptul că populațiile din care au fost extrase eșantioanele nu sunt omogene din punct de vedere a frecvențelor de apariție ale variabilei categoriale, iar în situația în care $\chi^2_{calculat} < \chi^2_{tabel,\ df}$ considerăm că nu avem suficiente dovezi pentru a respinge ipoteza nulă a unor populații omogene.

Scanned with OKEN Scanner

Testul χ^2 de verificare a omogenității pentru analiza a 2 eșantioane oferă o metodă alternativă a testului z, prezentat anterior în subcapitolul 9.5, în cazul testării ipotezei nule a egalității proporțiilor celor două populații din care provin eșantioanele, deoarece:

$$\chi^2_{df=1} = z^2 \tag{12.10}$$

Exemplu: în urma investigării unui eșantion de 180 de persoane (90 subiecți clinic sănătoși și 90 cu vicii de refracție) dorim să verificăm dacă în cele două populații de proveniență proporția persoanelor care prezintă cefalee diferă semnificativ. Datele obținute la nivelul eșantionul sunt prezintate în tabelul 12.8:

Tabelul 12.8. Tabelul cu datele inițiale ale celor 192 de persoane din eșantion

Pacient	**Variabila de grupare** (vicii de refracție)	**Variabila răspuns** (cefalee)
1	DA	DA
2	DA	DA
3	DA	NU
...	...	...
179	NU	NU
180	NU	DA

Se dorește studierea frecvenței de apariție a cefaleei în cazul a 2 eșantioane de persoane: 90 subiecți clinic sănătoși și 90 cu vicii de refracție. Cu alte cuvinte dorim să cunoaștem dacă cele două populații de persoane (cu vicii de refracție și sănătoase) din care provin aceste 2 eșantioane sunt omogene din punctul de vedere al frecvenței de apariție a cefaleei.

Vom realiza și în acest caz o centralizare a datelor după cele 2 variabile de interes într-un tabel de contingență (tabelul 12.9) și o reprezentare grafică a acestora (figura 12.3).

Tabelul 12.9. Frecvențele observate ale celor două variabile

Frecvențele observate	**Cefalee**		
Vicii de refracție	**DA**	**NU**	**TOTAL**
DA	26	64	**90**
NU	19	71	**90**
TOTAL	**45**	**135**	**180**

Scanned with OKEN Scanner

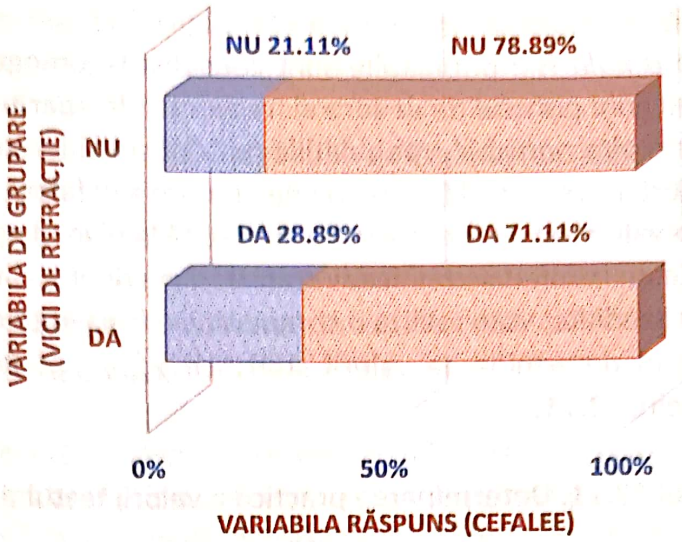


Figura. 12.3. Reprezentarea grafică a frecvențelor observate

Atât din analiza tabelului de contingență 12.9 cât și a graficului 12.3 remarcăm o pondere ușor mai ridicată (o frecvență mai crescută) în cazul persoanelor cu cefalee care aparțin eșantionului de persoane cu vicii de refracție (28,89% sau 26 persoane din totalul de 90) comparativ cu ponderea persoanelor cu cefalee din eșantionul de persoane clinic sănătoase (21,11% sau 19 persoane din totalul de 90). La prima vedere, se pare că cele două grupuri studiate nu sunt omogene din punctul de vedere al frecvenței cefaleei.

În continuare vom determina frecvențele așteptate corespunzătoare frecvențelor observate (tabelul 12.10), în condițiile unei ipoteze nule H_0 adevărate a omogenității celor două populații:

Tabelul 12.10. Frecvențele așteptate ale celor două variabile

Frecvențele așteptate	**Cefalee**		
Vicii de refracție	**DA**	**NU**	**TOTAL**
DA	22,5	67,5	**90,0**
NU	22,5	67,5	**90,0**
TOTAL	**45,0**	**135,0**	**180,0**

Având în vedere faptul că cele două eșantioane, de persoane sănătoase și persoane cu vicii de refracție, au același volum (90 de cazuri), rezultă imediat că frecvențele așteptate ale persoanelor cu cefalee în cazul când considerăm că populațiile de proveniență sunt omogene (ipoteza nulă H_0 adevărată) vor fi identice pentru fiecare în parte.

Scanned with OKEN Scanner

Ipotezele testului χ^2 de evaluare a omogenității acestor două populații sunt:

- *Ipoteza nulă* H_0: populațiile sunt considerate omogene (distribuite identic) din punctul de vedere al frecvenței de apariție a cefaleei;
- *Ipoteza alternativă* H_1: populațiile nu sunt considerate omogene din punctul de vedere al frecvenței de apariție a cefaleei.

Pentru a evalua șansa ca observația efectuată la nivelul eșantioanelor de lucru să reprezinte un rezultat semnificativ statistic și nu doar o simplă constatare la nivelul loturilor studiate, vom realiza o comparație a frecvențelor observate cu cele așteptate, prin determinarea valorii statisticii $\chi^2_{calculat}$. Rezultatele sunt prezentate în tabelul 12.11:

Tabelul 12.11. Determinarea practică a valorii testului χ^2

Test χ^2	Cefalee		
Vicii de refracție	**DA**	**NU**	**TOTAL**
DA	0,54	0,18	0,72
NU	0,54	0,18	0,72
TOTAL	1,08	0,36	**1,44**

Însumând cele patru valori determinate anterior: $0{,}54 + 0{,}18 + 0{,}54 + 0{,}18$ obținem valoarea calculată a testului: $\chi^2_{calculat} = 1{,}44$.

În continuare, vom citi din tabelele distribuției χ^2 valoarea critică, pentru un nivel de semnificație $\alpha = 0{,}05$ și un grad de libertate: $\chi^2_{tabel,1} = 3{,}841$. Deoarece $\chi^2_{calculat} = 1{,}44 < \chi^2_{tabel,1} = 3{,}841$, concluzionăm faptul că nu avem suficiente dovezi furnizate de eșantioanele studiate, pentru a putea respinge ipoteza nulă a omogenității celor două populații, din care au fost extrase în mod aleatoriu cele două eșantioane, cu privire la frecvența de apariție a cefaleei. De asemenea, valoarea probabilității asociate testului χ^2 este $p = 0{,}23$, cu mult superioară nivelului prag fixat de 0,05, aspect care confirmă riscul foarte ridicat de respingere a unei ipoteze nule adevărate. În final, concluzionăm că cele 2 populații pot fi considerate omogene în raport cu frecvența de apariție a cefaleei, diferențele evidențiate inițial, la nivel descriptiv, între cele două eșantioane putând fi astfel puse pe seama întâmplării.

12.4. RISCUL RELATIV UTILIZAT ÎN CADRUL STUDIILOR PROSPECTIVE

În cadrul studiilor *intervenționale*, așa cum sunt cele care se bazează pe analiza variației (ANOVA), cel puțin una dintre variabilele studiate este supusă unor manipulări, în concordanță cu designul specific experimentului. Pe lângă

Scanned with OKEN Scanner

acestea, în cadrul investigațiilor științifice, un loc important îl ocupă studiile *observaționale*, realizate sub forma evaluării și monitorizării stării de sănătate a pacienților, fără ca cercetătorul să intervină asupra expunerii.

Cea mai simplă formă pe care o poate îmbrăca un studiu observațional este cea în care se analizează 2 variabile de interes, una reprezentând factorul de risc (variabila independentă) și cealaltă variabila rezultat (variabila dependentă: prezența sau absența maladiei). De exemplu, variabila rezultat poate fi reprezentată de prezența sau absența cancerului pulmonar, iar variabila factor de risc poate fi exprimată prin statusul de fumător sau nefumător al pacienților din lotul studiat. Există două tipuri principale de studii observaționale: prospective și retrospective.

Studiile prospective reprezintă o categorie de studii observaționale în care sunt selectate în mod aleatoriu două eșantioane de persoane, unul care conține subiecții care prezintă factorul de risc și celălalt care conține subiecții care nu prezintă factorul de risc. Persoanele sunt urmărite prospectiv (în viitor) și este înregistrat numărul de subiecți din fiecare eșantion, care, în același moment de timp, se încadrează în fiecare categorie a variabilei rezultat.

Dacă variabilele studiate sunt binare, datele vor fi prezentate sub forma unui tabel de contingență (2x2):

Tabelul 12.12. Prezentarea eșantionului de pacienți în funcție de prezența maladiei și a factorului de risc

Frecvențele observate		**Prezența maladiei**		**Total**
		DA	**NU**	
Factor de risc	**DA**	a	b	$\boldsymbol{a+b}$
	NU	c	d	$\boldsymbol{c+d}$
Total		$\boldsymbol{a+c}$	$\boldsymbol{b+d}$	$\boldsymbol{a+b+c+d=n}$

În cazul studiilor prospective, se cunoaște că întâi a fost factorul de risc prezent și apoi a apărut boala, direcția cercetării fiind de la expunere către efect

Riscul reprezintă probabilitatea ca un eveniment de interes (o maladie, o complicație sau chiar decesul) să apară într-o anumită populație într-un interval de timp specific, în asociere cu o expunere sau un factor de risc anume.

În cadrul studiilor prospective vom determina două categorii de riscuri:

a) riscul de a dezvolta o maladie în cazul pacienților expuși la acțiunea factorului de risc:

$$Risc_{expusi} = \frac{Pacienti\ expusi\ care\ au\ dezvoltat\ maladia}{Total\ pacienti\ expusi\ factorului\ de\ risc} = \frac{a}{a+b} \quad (12.11)$$

Scanned with OKEN Scanner

b) riscul de a dezvolta o maladie în cazul pacienților care nu au fost expuși la acțiunea factorului de risc:

$$Risc_{neexpusi} = \frac{Pacienti\ neexpusi\ care\ au\ dezvoltat\ maladia}{Total\ pacienti\ neexpusi\ factorului\ de\ risc} = \frac{c}{c+d} \quad (12.12)$$

Riscul relativ (RR) se utilizează în cazul studiilor prospective, în care persoanele care sunt supuse sau nu unui factor de risc sunt urmărite în timp, reprezentând o măsură a asocierii între factorul de risc și afecțiunea studiată. Indicatorul se calculează ca raport între frecvența de apariție a afecțiunii în cadrul populației supusă factorului de risc și frecvența de apariție a maladiei în populația în care factorul de risc nu este prezent (relația 12.13). Cu alte cuvinte, riscul relativ compară probabilitatea de a dezvolta o maladie la un grup expus factorului de risc cu probabilitatea de a dezvolta maladia la un grup care nu a fost expus:

$$\widehat{RR} = \frac{Risc_{expusi}}{Risc_{neexpusi}} = \frac{a/a+b}{c/c+d} = \frac{a \cdot (c+d)}{c \cdot (a+b)} \quad (12.13)$$

$\widehat{RR}$ din relația 12.13 reprezintă riscul relativ calculat pe baza datelor din eșantionul de lucru, care este utilizat cu rolul de estimator al riscului relativ RR din populația generală, din care eșantionul a fost extras în mod aleatoriu.

Domeniul de variație al riscului relativ este cuprins între 0 și $+\infty$.

Deosebim 3 situații:

- $RR = 1$, cu alte cuvinte factorul de risc nu are niciun efect asupra frecventei de apariție a maladiei;
- $RR > 1$ indică o creștere a frecvenței maladiei pentru cazurile cu factor prezent, ceea ce evidențiază un efect negativ;
- $RR < 1$ indică un efect pozitiv (de protecție) al factorului de risc, acesta ducând la scăderea frecvenței de apariție a afecțiunii.

Valoarea 1 este considerată de referință, deoarece în funcție de aceasta se deduce semnificația statistică a potențialei asocieri între variabile.

Riscul relativ calculat la nivel de eșantion $\widehat{RR}$ reprezintă o estimare punctuală a parametrului corespunzător al populației din care acesta provine RR. Se poate construi un interval de încredere 95% pentru valoarea „necunoscută" și „adevărată" a riscului relativ al populației generale, din care eșantionul fost extras.

Pentru construirea intervalului de încredere 95% al riscului relativ din populația generală, se folosește transformarea logaritmică, deoarece distribuția lui $ln(RR)$ este mai simetrică, având o alură aproximativ normală.

Scanned with OKEN Scanner

Logaritmul natural din raportul riscurilor are o distribuție normală standard, de medie zero și varianță 1:

$$ln(\widehat{RR}) = ln\frac{a/a+b}{c/c+d} \rightarrow N(0,1) \quad (12.14)$$

Eroarea standard a estimării logaritmului natural din raportul riscurilor $ln(\widehat{RR})$ este:

$$SE_{ln\,(RR)} = \sqrt{\frac{b}{a\cdot(a+b)} + \frac{d}{c\cdot(c+d)}} \quad (12.15)$$

Intervalul de încredere estimat cu o probabilitate de 95% pentru logaritmul natural al raportului riscurilor din populația generală $ln(RR)$ va fi:

$$ln(\widehat{RR}) - z_{1-\frac{\alpha}{2}} \cdot SE_{ln\,(RR)} < ln\,(RR) < ln(\widehat{RR}) + z_{1-\frac{\alpha}{2}} \cdot SE_{ln\,(RR)} \quad (12.16)$$

În final, pentru a reveni la raportul riscurilor RR, vom exponența cele două limite ale intervalului din relația (12.16):

$$e^{ln(\widehat{RR}) - z_{1-\frac{\alpha}{2}} \cdot SE_{ln\,(RR)}} < RR < e^{ln(\widehat{RR}) + z_{1-\frac{\alpha}{2}} \cdot SE_{ln\,(RR)}} \quad (12.17)$$

În situația în care intervalul de confidență conține valoarea 1, rezultă că factorul de risc nu influențează frecvența de apariție a maladiei în cadrul populației studiate ($Risc_{expusi} = Risc_{neexpusi}$).

Exemplu: revenim datele privind posibila asociere între factorul de risc - statusul (fumător / nefumător) și maladie - prezenta cancerului pulmonar, pentru eșantionul de 100 de persoane prezentat în subcapitolul 12.2.

Vom calcula riscul de a dezvolta cancer pulmonar pentru persoanele expuse la acțiunea factorului de risc (fumători) precum și pentru persoanele care nu au fost expuse (nefumători). Pentru calculul indicatorilor, folosim datele centralizate în tabelul 12.5 și aplicăm formulele (12.11) și (12.12):

$$Risc_{expusi} = \frac{Pacienti\ fumatori\ care\ au\ dezvoltat\ cancer}{Total\ pacienti\ fumatori} = \frac{38}{53} = 0{,}717$$

$$Risc_{neexpusi} = \frac{Pacienti\ nefumatori\ care\ au\ dezvoltat\ cancer}{Total\ pacienti\ nefumatori} = \frac{14}{47} = 0{,}298$$

Scanned with OKEN Scanner

Prin raportarea celor două riscuri determinate anterior, vom obține riscul relativ calculat pe baza datelor din eșantion, în calitate de estimator punctual al riscului relativ RR din populația generală (relația 12.13):

$$\widehat{RR} = \frac{Risc_{expusi}}{Risc_{neexpusi}} = \frac{0,717}{0,298} = 2,407$$

Prin urmare, datele din eșantionul supus investigației evidențiază un risc de a dezvolta cancer pulmonar în cazul pacienților expuși factorului de risc (fumători) de 2,407 ori mai mare comparativ cu pacienții care nu au fost expuși (nefumători). Trebuie ținut însă cont de faptul că $\widehat{RR} = 2,407$ reprezintă doar o estimație punctuală pentru valoarea riscului relativ de la nivelul populației generale RR, lotul analizat reprezentând doar un singur eșantion din numărul enorm de eșantioane de același volum care ar putea fi extrase în mod aleatoriu din această populație. Din acest considerent, este necesară construirea unui interval de încredere estimat cu o probabilitate de 95% pentru RR din populația generală.

Calculăm eroarea standard a estimării logaritmului natural din raportul riscurilor $ln(\widehat{RR})$ utilizând relația (12.15):

$$SE_{ln\,(RR)} = \sqrt{\frac{15}{38 \cdot (38+15)} + \frac{33}{14 \cdot (14+33)}} = \sqrt{0,0576} = 0,24$$

În final, construim intervalul de încredere 95% pentru RR al populației cu formula (12.17). Pentru determinarea valorii critice $z_{1-\alpha/2}$ putem folosi tabelele distribuției normale standard, pentru un nivel de încredere de 95% ($1 - \alpha = 0,95$ și $\alpha = 0,05$): $z_{1-\alpha/2} = 1,96$. Astfel, avem:

$$e^{ln(2,41)-1,96 \cdot 0,24} < RR < e^{ln(2,41)-1,96 \cdot 0,24}$$

și, efectuând calculele obținem: $1,504 < RR < 3,853$

Prin urmare avem 95% încredere că raportul riscurilor de a dezvolta cancer pulmonar din populația generală în cazul persoanelor fumătoare comparativ cu cele nefumătoare se situează în intervalul cuprins între 1,504 și 3,853. Deoarece intervalul de încredere determinat nu include valoarea 1, concluzionăm că, la nivel populațional, există un risc mai mare ca persoanele fumătoare să dezvolte cancer pulmonar și în consecință putem afirma că asocierea dintre factorul de risc (fumatul) și probabilitatea de a dezvolta maladia (cancerul pulmonar) este semnificativă statistic.

Scanned with OKEN Scanner

12.5. RAPORTUL COTELOR UTILIZAT ÎN CADRUL STUDIILOR RETROSPECTIVE

Studiile retrospective reprezintă varianta opusă a studiilor prospective, eșantioanele fiind selectate în prezent din populațiile categoriilor variabilei rezultat (maladie prezentă sau absentă). Cercetătorul privește retrospectiv subiecții selectați (în trecut) și determină la care dintre aceștia factorul de risc a fost prezent sau absent. Datele vor fi sistematizate sub forma unui tabel de contingență similar cu cel din situația studiilor prospective.

În situația studiilor de tip caz-martor (case-control), pentru a exprima legătura existentă, se folosește indicatorul *raportul cotelor* sau *raportul șanselor* (OR). Studiul de acest tip este retrospectiv, având ca scop realizarea unei comparații între persoanele care prezintă maladia (cazuri), respectiv cele ce nu prezintă maladia (martori sau control), grupate după prezența sau nu a factorului de risc.

Cota, utilizată în mod frecvent de jucătorii care preferă pariurile sportive, se definește ca fiind raportul dintre probabilitatea de apariție a evenimentului studiat (succes) și probabilitatea de realizare a evenimentului complementar sau opus (insucces).

Datele vor fi sistematizate într-un tabel relativ similar cu cel prezentat în cazul studiilor prospective:

Tabelul 12.13. Prezentarea eșantionului de pacienți în funcție de prezența maladiei și a factorului de risc

Frecvențele observate		**Eșantioane**		**Total**
		Cazuri (au maladia)	**Control (nu au maladia)**	
Factor de risc	**Prezent**	a	b	$\boldsymbol{a+b}$
	Absent	c	d	$\boldsymbol{c+d}$
Total		$\boldsymbol{a+c}$	$\boldsymbol{b+d}$	$\boldsymbol{n}$

În cadrul studiilor retrospective se pot calcula 2 tipuri de cote:

- șansa de a fi un caz față de a fi un control printre subiecții având *factorul de risc prezent*:

$$Sansa_{caz/control}^{risc\ prezent} = \frac{Probabilitatea_{caz/risc\ prezent}}{Probabilitatea_{control/risc\ prezent}} = \frac{a/_{a+b}}{b/_{a+b}} = \frac{a}{b} \quad (12.18)$$

Scanned with OKEN Scanner

- șansa de a fi un caz față de a fi un control printre subiecții având *factorul de risc absent*:

$$Sansa_{caz/control}^{risc\ absent} = \frac{Probabilitatea_{caz/risc\ absent}}{Probabilitatea_{control/risc\ absent}} = \frac{c/c+d}{d/c+d} = \frac{c}{d} \tag{12.19}$$

Raportul acestor două cote (raportul șanselor $\widehat{OR}$) pentru un studiu retrospectiv, calculat pe baza datelor din eșantion, va fi utilizat ca estimator punctual al raportului șanselor din populația generală OR:

$$\widehat{OR} = \frac{Sansa_{caz/control}^{risc\ prezent}}{Sansa_{caz/control}^{risc\ absent}} = \frac{a/b}{c/d} = \frac{a \cdot d}{b \cdot c} \tag{12.20}$$

Domeniul de variație al raportului cotelor este cuprins între 0 și $+\infty$.

Și în acest caz putem determina intervalul de încredere 95% pentru raportul șanselor OR din populația generală.

Logaritmul natural din raportul șanselor urmează o distribuție normală standard:

$$ln\left(\widehat{OR}\right) = ln\frac{a \cdot d}{b \cdot c} \rightarrow N(0,1) \tag{12.21}$$

Eroarea standard a logaritmului natural din raportul șanselor este:

$$SE_{ln\,(OR)} = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \tag{12.22}$$

Intervalul de încredere estimat cu o probabilitate de 95% pentru logaritmul natural al raportului cotelor din populația generală $ln(OR)$ va fi:

$$ln\left(\widehat{OR}\right) - z_{1-\frac{\alpha}{2}} \cdot SE_{\ln(OR)} < ln(OR) < ln\left(\widehat{OR}\right) + z_{1-\frac{\alpha}{2}} \cdot SE_{\ln(OR)} \tag{12.23}$$

Ulterior, vom exponența, obținând în acest fel intervalul de încredere 95% al raportului cotelor din populația generală OR:

$$e^{ln(\widehat{OR}) - z_{1-\frac{\alpha}{2}} \cdot SE_{\ln(\widehat{OR})}} < OR < e^{ln(\widehat{OR}) + z_{1-\frac{\alpha}{2}} \cdot SE_{ln(\widehat{OR})}} \tag{12.24}$$

Dacă intervalul de confidență al raportului cotelor conține valoarea 1, rezultă că nu există nicio asociere între factorul de risc și prezența sau absența maladiei.

Scanned with OKEN Scanner

Interpretare: șansele de a se îmbolnăvi sunt de OR ori mai mari la persoanele care au prezent factorul de risc comparativ cu cei la care acesta este absent.

Atât riscul relativ cât și raportul șanselor exprimă în esență aceeași caracteristică a datelor, fiind practic 2 indicatori care cuantifică în mod proporțional intensitatea asocierii existente. În cazul unei afecțiuni sau maladii rare, raportul cotelor din populația studiată oferă o bună aproximare a riscului relativ din cadrul aceleiași populații.

Exemplu: vom utiliza aceleași date prezentate anterior privind posibila asociere între factorul de risc - statusul (fumător / nefumător) și maladie - prezenta cancerului pulmonar, pentru eșantionul de 100 de persoane, privit de data aceasta într-o manieră retrospectivă. Prin urmare, dorim să comparăm șansele de a avea cancer pulmonar pentru subiecții care au fost fumători (factor de risc prezent) cu șansele de a avea maladia în cazul subiecților care nu au fost fumători (factor de risc absent).

Vom calcula șansele de a avea cancer pulmonar pentru subiecții care au fost fumători precum și pentru subiecții care nu au fost fumători. Pentru calculul indicatorilor aplicăm formulele (12.18) și (12.19) utilizând datele din tabelul 12.5:

$$Sansa\ fumatori_{caz/control}^{risc\ prezent} = \frac{38/53}{15/53} = \frac{38}{15} = 2{,}533$$

$$Sansa\ nefumatori_{caz/control}^{risc\ absent} = \frac{14/47}{33/47} = \frac{14}{33} = 0{,}424$$

Prin raportarea celor două șanse (cote) determinate anterior, vom obține raportul cotelor calculat pe baza datelor din eșantion, în calitate de estimator punctual al raportului cotelor OR din populația generală (relația 12.20):

$$\widehat{OR} = \frac{Sansa\ fumatori_{caz/control}^{risc\ prezent}}{Sansa\ nefumatori_{caz/control}^{risc\ absent}} = \frac{2{,}533}{0{,}424} = 5{,}971$$

Potrivit datelor din eșantionul investigat, în cazul subiecților care au fost fumători șansele de a avea cancer pulmonar sunt de aproximativ 6 ori mai mari comparativ cu șansele persoanelor care au fost nefumătoare.

Vom proceda și în acest caz la construirea unui interval de încredere estimat cu o probabilitate de 95% pentru OR al populației generale. În acest sens, calculăm eroarea standard a estimării logaritmului natural din raportul cotelor $ln(\widehat{OR})$ utilizând relația (12.15):

Scanned with OKEN Scanner

$$SE_{ln\,(OR)} = \sqrt{\frac{1}{38} + \frac{1}{15} + \frac{1}{14} + \frac{1}{33}} = 0{,}441$$

În etapa finală, construim intervalul de încredere 95% pentru OR al populației generale cu formula (12.24). Pentru determinarea valorii critice $z_{1-\alpha/2}$ putem folosi tabelele distribuției normale standard, pentru un nivel de încredere de 95% ($1 - \alpha = 0{,}95$ și $\alpha = 0{,}05$): $z_{1-\alpha/2} = 1{,}96$. Astfel, avem:

$$e^{ln(5{,}971) - 1{,}96 \cdot 0{,}441} < OR < e^{ln(5{,}971) + 1{,}96 \cdot 0{,}441}$$

și, efectuând calculele obținem: $2{,}515 < OR < 14{,}180$

Prin urmare avem 95% încredere că raportul șanselor de a avea cancer pulmonar din populația generală în cazul persoanelor care au fost fumătoare comparativ cu cele care au fost nefumătoare se situează în intervalul cuprins între 2,515 și 14,180. Cu alte cuvinte, persoanele expuse prezintă de 2,515 până la 14,180 ori șanse mai mari de a dezvolta maladia comparativ cu cele neexpuse. Deoarece intervalul de încredere calculat anterior nu include valoarea 1, concluzionăm că, la nivel populațional, este mai probabil ca persoanele cu cancer pulmonar să fi fost fumătoare în comparație cu cele care nu au avut cancer pulmonar.

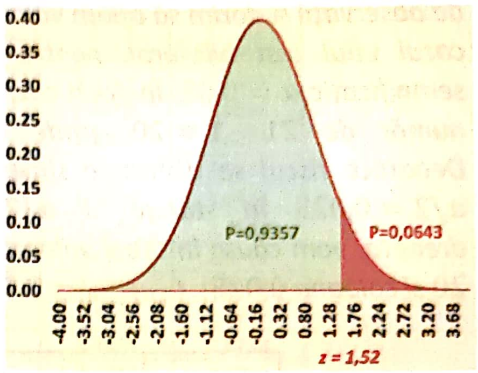
Scanned with OKEN Scanner

ANEXE

Toate valorile din prezentul capitol sunt calculate de autor, utilizând programul Excel

1. Distribuția normală standard (z)

Pentru a citi o valoare din tabel, localizăm, de exemplu, $z = 1{,}52$ în tabelul distribuției normale standard și preluăm nivelul corespunzător al probabilității cumulate de 0,9357, situat la intersecția liniei 1,50 și a coloanei cu a doua zecimală 0,02. Valoarea de 0,9357 sau 93,57% reprezintă probabilitatea ca o un anumit z extras în mod aleatoriu din populația Z să aibă o valoare cuprinsă între $-\infty$ și 1,52.



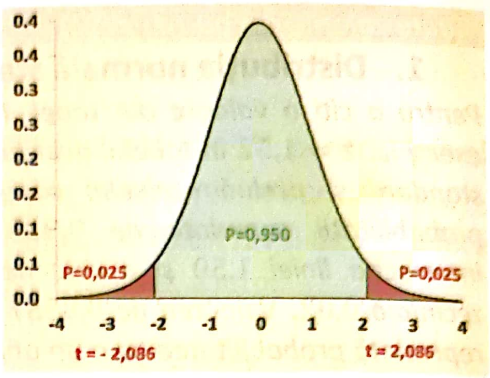
Dacă dorim să aflăm probabilitatea ca o variabilă statistică, care urmează o distribuție normală standard, să fie mai mare decât 1,96 vom determina inițial $P(z < 1{,}96)$, care conform tabelului este 0,9750 și apoi vom scădea rezultatul din 1: $P(z > 1{,}96) = 1 - P(z < 1{,}96) = 1 - 0{,}9750 = 0{,}025$ sau 2,5%.

z	Precizia									
	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,10	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,20	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,30	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,40	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,50	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,60	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,70	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,80	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,90	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,00	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,10	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,20	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,30	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,40	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,50	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,60	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,70	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,80	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,90	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,00	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,10	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,20	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,30	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,40	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,50	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,60	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,70	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,80	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,90	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,00	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990

Scanned with OKEN Scanner

2. Distribuția Student (t)

Presupunem că avem un eșantion format din 21 de observații și dorim să aflăm valoarea critică în cazul unui test bilateral, pentru un nivel de semnificație $\alpha = 0{,}05$. În acest caz, vom avea un număr de $21 - 1 = 20$ grade de libertate. Deoarece riscul se distribuie simetric bilateral, $\alpha/2 = 0{,}025$ în stânga și $\alpha/2 = 0{,}025$ în dreapta, vom căuta în tabel valoarea de pe linia 20 și coloana 0,025: $t_{0{,}025;20} = 2{,}0860$



Grade de libertate	Suprafața delimitată în extremitatea dreaptă				
	0,10	0,05	0,025	0,01	0,005
1	3,0777	6,3138	12,7062	31,8205	63,6567
2	1,8856	2,9200	4,3027	6,9646	9,9248
3	1,6377	2,3534	3,1824	4,5407	5,8409
4	1,5332	2,1318	2,7764	3,7469	4,6041
5	1,4759	2,0150	2,5706	3,3649	4,0321
6	1,4398	1,9432	2,4469	3,1427	3,7074
7	1,4149	1,8946	2,3646	2,9980	3,4995
8	1,3968	1,8595	2,3060	2,8965	3,3554
9	1,3830	1,8331	2,2622	2,8214	3,2498
10	1,3722	1,8125	2,2281	2,7638	3,1693
11	1,3634	1,7959	2,2010	2,7181	3,1058
12	1,3562	1,7823	2,1788	2,6810	3,0545
13	1,3502	1,7709	2,1604	2,6503	3,0123
14	1,3450	1,7613	2,1448	2,6245	2,9768
15	1,3406	1,7531	2,1314	2,6025	2,9467
16	1,3368	1,7459	2,1199	2,5835	2,9208
17	1,3334	1,7396	2,1098	2,5669	2,8982
18	1,3304	1,7341	2,1009	2,5524	2,8784
19	1,3277	1,7291	2,0930	2,5395	2,8609
20	1,3253	1,7247	2,0860	2,5280	2,8453
21	1,3232	1,7207	2,0796	2,5176	2,8314
22	1,3212	1,7171	2,0739	2,5083	2,8188
23	1,3195	1,7139	2,0687	2,4999	2,8073
24	1,3178	1,7109	2,0639	2,4922	2,7969
25	1,3163	1,7081	2,0595	2,4851	2,7874
26	1,3150	1,7056	2,0555	2,4786	2,7787
27	1,3137	1,7033	2,0518	2,4727	2,7707
28	1,3125	1,7011	2,0484	2,4671	2,7633
29	1,3114	1,6991	2,0452	2,4620	2,7564
30	1,3104	1,6973	2,0423	2,4573	2,7500
31	1,3095	1,6955	2,0395	2,4528	2,7440
32	1,3086	1,6939	2,0369	2,4487	2,7385
33	1,3077	1,6924	2,0345	2,4448	2,7333
34	1,3070	1,6909	2,0322	2,4411	2,7284
35	1,3062	1,6896	2,0301	2,4377	2,7238
36	1,3055	1,6883	2,0281	2,4345	2,7195
37	1,3049	1,6871	2,0262	2,4314	2,7154
38	1,3042	1,6860	2,0244	2,4286	2,7116
39	1,3036	1,6849	2,0227	2,4258	2,7079
40	1,3031	1,6839	2,0211	2,4233	2,7045
41	1,3025	1,6829	2,0195	2,4208	2,7012
42	1,3020	1,6820	2,0181	2,4185	2,6981

Scanned with OKEN Scanner

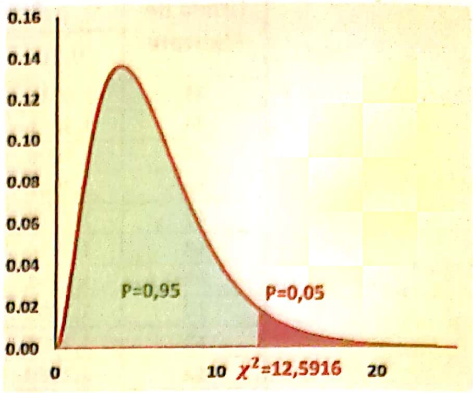
continuarea tabelului distribuției Student

Grade de libertate	Suprafața delimitată în extremitatea dreaptă				
	0,10	0,05	0,025	0,01	0,005
43	1,3016	1,6811	2,0167	2,4163	2,6951
44	1,3011	1,6802	2,0154	2,4141	2,6923
45	1,3006	1,6794	2,0141	2,4121	2,6896
46	1,3002	1,6787	2,0129	2,4102	2,6870
47	1,2998	1,6779	2,0117	2,4083	2,6846
48	1,2994	1,6772	2,0106	2,4066	2,6822
49	1,2991	1,6766	2,0096	2,4049	2,6800
50	1,2987	1,6759	2,0086	2,4033	2,6778
51	1,2984	1,6753	2,0076	2,4017	2,6757
52	1,2980	1,6747	2,0066	2,4002	2,6737
53	1,2977	1,6741	2,0057	2,3988	2,6718
54	1,2974	1,6736	2,0049	2,3974	2,6700
55	1,2971	1,6730	2,0040	2,3961	2,6682
56	1,2969	1,6725	2,0032	2,3948	2,6665
57	1,2966	1,6720	2,0025	2,3936	2,6649
58	1,2963	1,6716	2,0017	2,3924	2,6633
59	1,2961	1,6711	2,0010	2,3912	2,6618
60	1,2958	1,6706	2,0003	2,3901	2,6603
61	1,2956	1,6702	1,9996	2,3890	2,6589
62	1,2954	1,6698	1,9990	2,3880	2,6575
63	1,2951	1,6694	1,9983	2,3870	2,6561
64	1,2949	1,6690	1,9977	2,3860	2,6549
65	1,2947	1,6686	1,9971	2,3851	2,6536
66	1,2945	1,6683	1,9966	2,3842	2,6524
67	1,2943	1,6679	1,9960	2,3833	2,6512
68	1,2941	1,6676	1,9955	2,3824	2,6501
69	1,2939	1,6672	1,9949	2,3816	2,6490
70	1,2938	1,6669	1,9944	2,3808	2,6479
71	1,2936	1,6666	1,9939	2,3800	2,6469
72	1,2934	1,6663	1,9935	2,3793	2,6459
73	1,2933	1,6660	1,9930	2,3785	2,6449
74	1,2931	1,6657	1,9925	2,3778	2,6439
75	1,2929	1,6654	1,9921	2,3771	2,6430
76	1,2928	1,6652	1,9917	2,3764	2,6421
77	1,2926	1,6649	1,9913	2,3758	2,6412
78	1,2925	1,6646	1,9908	2,3751	2,6403
79	1,2924	1,6644	1,9905	2,3745	2,6395
80	1,2922	1,6641	1,9901	2,3739	2,6387
81	1,2921	1,6639	1,9897	2,3733	2,6379
82	1,2920	1,6636	1,9893	2,3727	2,6371
83	1,2918	1,6634	1,9890	2,3721	2,6364
84	1,2917	1,6632	1,9886	2,3716	2,6356
85	1,2916	1,6630	1,9883	2,3710	2,6349
86	1,2915	1,6628	1,9879	2,3705	2,6342
87	1,2914	1,6626	1,9876	2,3700	2,6335
88	1,2912	1,6624	1,9873	2,3695	2,6329
89	1,2911	1,6622	1,9870	2,3690	2,6322
90	1,2910	1,6620	1,9867	2,3685	2,6316
91	1,2909	1,6618	1,9864	2,3680	2,6309
92	1,2908	1,6616	1,9861	2,3676	2,6303
93	1,2907	1,6614	1,9858	2,3671	2,6297
94	1,2906	1,6612	1,9855	2,3667	2,6291
95	1,2905	1,6611	1,9853	2,3662	2,6286
96	1,2904	1,6609	1,9850	2,3658	2,6280
97	1,2903	1,6607	1,9847	2,3654	2,6275
98	1,2902	1,6606	1,9845	2,3650	2,6269
99	1,2902	1,6604	1,9842	2,3646	2,6264
100	1,2901	1,6602	1,9840	2,3642	2,6259

Scanned with OKEN Scanner

3. Distribuția Chi-pătrat (χ^2)

Presupunem că în cadrul unui test χ^2 de evaluare a independenței dintre două variabile categoriale, una având 3 și cealaltă 4 categorii, utilizăm un nivel de semnificație $\alpha = 0{,}05$. Vom avea $(3-1)\cdot(4-1) = 2\cdot 3 = 6$ grade de libertate. Prin urmare, valoarea critică din tabelul distribuției χ^2 va fi situată la intersecția liniei 6 cu coloana 0,05: $\chi^2_{0{,}05;6} = 12{,}5916$ (riscul este distribuit unilateral în partea dreaptă a distribuției).



Grade de libertate	Suprafața delimitată în extremitatea dreaptă			
	0,100	0,050	0,010	0,001
1	2,7055	3,8415	6,6349	10,8276
2	4,6052	5,9915	9,2103	13,8155
3	6,2514	7,8147	11,3449	16,2662
4	7,7794	9,4877	13,2767	18,4668
5	9,2364	11,0705	15,0863	20,5150
6	10,6446	12,5916	16,8119	22,4577
7	12,0170	14,0671	18,4753	24,3219
8	13,3616	15,5073	20,0902	26,1245
9	14,6837	16,9190	21,6660	27,8772
10	15,9872	18,3070	23,2093	29,5883
11	17,2750	19,6751	24,7250	31,2641
12	18,5493	21,0261	26,2170	32,9095
13	19,8119	22,3620	27,6882	34,5282
14	21,0641	23,6848	29,1412	36,1233
15	22,3071	24,9958	30,5779	37,6973
16	23,5418	26,2962	31,9999	39,2524
17	24,7690	27,5871	33,4087	40,7902
18	25,9894	28,8693	34,8053	42,3124
19	27,2036	30,1435	36,1909	43,8202
20	28,4120	31,4104	37,5662	45,3147
21	29,6151	32,6706	38,9322	46,7970
22	30,8133	33,9244	40,2894	48,2679
23	32,0069	35,1725	41,6384	49,7282
24	33,1962	36,4150	42,9798	51,1786
25	34,3816	37,6525	44,3141	52,6197
26	35,5632	38,8851	45,6417	54,0520
27	36,7412	40,1133	46,9629	55,4760
28	37,9159	41,3371	48,2782	56,8923
29	39,0875	42,5570	49,5879	58,3012
30	40,2560	43,7730	50,8922	59,7031
31	41,4217	44,9853	52,1914	61,0983
32	42,5847	46,1943	53,4858	62,4872
33	43,7452	47,3999	54,7755	63,8701
34	44,9032	48,6024	56,0609	65,2472
35	46,0588	49,8018	57,3421	66,6188
36	47,2122	50,9985	58,6192	67,9852
37	48,3634	52,1923	59,8925	69,3465
38	49,5126	53,3835	61,1621	70,7029
39	50,6598	54,5722	62,4281	72,0547
40	51,8051	55,7585	63,6907	73,4020
41	52,9485	56,9424	64,9501	74,7449
42	54,0902	58,1240	66,2062	76,0838

Scanned with OKEN Scanner

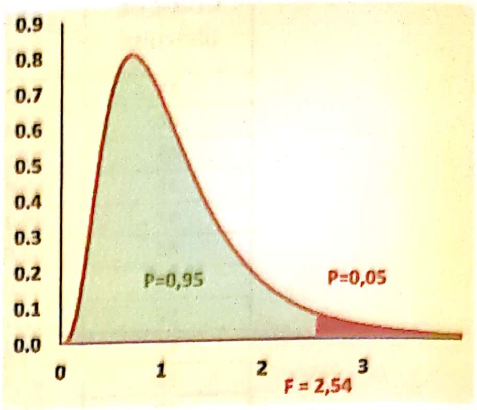
continuarea tabelului distribuției Chi-pătrat

Grade de libertate	Suprafața delimitată în extremitatea dreaptă			
	0,100	0,050	0,010	0,001
43	55,2302	59,3035	67,4593	77,4186
44	56,3685	60,4809	68,7095	78,7495
45	57,5053	61,6562	69,9568	80,0767
46	58,6405	62,8296	71,2014	81,4003
47	59,7743	64,0011	72,4433	82,7204
48	60,9066	65,1708	73,6826	84,0371
49	62,0375	66,3386	74,9195	85,3506
50	63,1671	67,5048	76,1539	86,6608
51	64,2954	68,6693	77,3860	87,9680
52	65,4224	69,8322	78,6158	89,2722
53	66,5482	70,9935	79,8433	90,5734
54	67,6728	72,1532	81,0688	91,8718
55	68,7962	73,3115	82,2921	93,1675
56	69,9185	74,4683	83,5134	94,4605
57	71,0397	75,6237	84,7328	95,7510
58	72,1598	76,7778	85,9502	97,0388
59	73,2789	77,9305	87,1657	98,3242
60	74,3970	79,0819	88,3794	99,6072
61	75,5141	80,2321	89,5913	100,8879
62	76,6302	81,3810	90,8015	102,1662
63	77,7454	82,5287	92,0100	103,4424
64	78,8596	83,6753	93,2169	104,7163
65	79,9730	84,8206	94,4221	105,9881
66	81,0855	85,9649	95,6257	107,2579
67	82,1971	87,1081	96,8278	108,5256
68	83,3079	88,2502	98,0284	109,7913
69	84,4179	89,3912	99,2275	111,0551
70	85,5270	90,5312	100,4252	112,3169
71	86,6354	91,6702	101,6214	113,5769
72	87,7430	92,8083	102,8163	114,8351
73	88,8499	93,9453	104,0098	116,0915
74	89,9560	95,0815	105,2020	117,3462
75	91,0615	96,2167	106,3929	118,5991
76	92,1662	97,3510	107,5825	119,8503
77	93,2702	98,4844	108,7709	121,1000
78	94,3735	99,6169	109,9581	122,3480
79	95,4762	100,7486	111,1440	123,5944
80	96,5782	101,8795	112,3288	124,8392
81	97,6796	103,0095	113,5124	126,0826
82	98,7803	104,1387	114,6949	127,3244
83	99,8805	105,2672	115,8763	128,5648
84	100,9800	106,3948	117,0565	129,8037
85	102,0789	107,5217	118,2357	131,0412
86	103,1773	108,6479	119,4139	132,2773
87	104,2750	109,7733	120,5910	133,5121
88	105,3722	110,8980	121,7671	134,7455
89	106,4689	112,0220	122,9422	135,9776
90	107,5650	113,1453	124,1163	137,2084
91	108,6606	114,2679	125,2895	138,4379
92	109,7556	115,3898	126,4617	139,6661
93	110,8502	116,5110	127,6329	140,8931
94	111,9442	117,6317	128,8032	142,1189
95	113,0377	118,7516	129,9727	143,3435
96	114,1307	119,8709	131,1412	144,5670
97	115,2232	120,9896	132,3089	145,7892
98	116,3153	122,1077	133,4757	147,0104
99	117,4069	123,2252	134,6416	148,2304
100	118,4980	124,3421	135,8067	149,4493

Scanned with OKEN Scanner

4. Distribuția Fisher (F)

Presupunem faptul că avem două eșantioane ($n_1 = 11$ și $n_2 = 16$) și dorim să testăm dacă variațele populațiilor din care acestea provin sunt diferite, la un nivel de semnificație $\alpha = 0{,}05$. Vom avea la numărător un număr de $11 - 1 = 10$ grade de libertate, iar la numitor $16 - 1 = 15$ grade de libertate (test F unilateral, dreapta). Citim în tabelul distribuției Fisher valoarea aflată la intersecția coloanei cu 10 grade de libertate și a liniei cu 15 grade de libertate: $F_{0,05;\,10;\,15} = 2{,}54$



Grade de libertate numitor	Grade de libertare numărător (1 - 12) ($\alpha = 0{,}05$)											
	1	2	3	4	5	6	7	8	9	10	11	12
1	161,4	199,5	215,7	224,6	230,2	234	236,8	238,9	240,5	241,9	243	243,9
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,40	19,41
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,76	8,74
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,94	5,91
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70	4,68
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60	3,57
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31	3,28
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10	3,07
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82	2,79
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72	2,69
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63	2,60
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,57	2,53
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,51	2,48
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,46	2,42
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,41	2,38
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37	2,34
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,34	2,31
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,31	2,28
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,28	2,25
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,26	2,23
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,24	2,20
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,22	2,18
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,20	2,16
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,18	2,15
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,17	2,13
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,15	2,12
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,14	2,10
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,13	2,09
31	4,16	3,30	2,91	2,68	2,52	2,41	2,32	2,25	2,20	2,15	2,11	2,08
32	4,15	3,29	2,90	2,67	2,51	2,40	2,31	2,24	2,19	2,14	2,10	2,07
33	4,14	3,28	2,89	2,66	2,50	2,39	2,30	2,23	2,18	2,13	2,09	2,06
34	4,13	3,28	2,88	2,65	2,49	2,38	2,29	2,23	2,17	2,12	2,08	2,05
35	4,12	3,27	2,87	2,64	2,49	2,37	2,29	2,22	2,16	2,11	2,07	2,04
36	4,11	3,26	2,87	2,63	2,48	2,36	2,28	2,21	2,15	2,11	2,07	2,03
37	4,11	3,25	2,86	2,63	2,47	2,36	2,27	2,20	2,14	2,10	2,06	2,02
38	4,10	3,24	2,85	2,62	2,46	2,35	2,26	2,19	2,14	2,09	2,05	2,02
39	4,09	3,24	2,85	2,61	2,46	2,34	2,26	2,19	2,13	2,08	2,04	2,01
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,04	2,00
41	4,08	3,23	2,83	2,60	2,44	2,33	2,24	2,17	2,12	2,07	2,03	2,00
42	4,07	3,22	2,83	2,59	2,44	2,32	2,24	2,17	2,11	2,06	2,03	1,99

Scanned with OKEN Scanner

continuarea tabelului distribuției Fisher

Grade de libertate numitor	Grade de libertare numărător (13 - 24) ($\alpha = 0,05$)											
	13	14	15	16	17	18	19	20	21	22	23	24
1	244,7	245,4	245,9	246,5	246,9	247,3	247,7	248	248,3	248,6	248,8	249,1
2	19,42	19,42	19,43	19,43	19,44	19,44	19,44	19,45	19,45	19,45	19,45	19,45
3	8,73	8,71	8,70	8,69	8,68	8,67	8,67	8,66	8,65	8,65	8,64	8,64
4	5,89	5,87	5,86	5,84	5,83	5,82	5,81	5,80	5,79	5,79	5,78	5,77
5	4,66	4,64	4,62	4,60	4,59	4,58	4,57	4,56	4,55	4,54	4,53	4,53
6	3,98	3,96	3,94	3,92	3,91	3,90	3,88	3,87	3,86	3,86	3,85	3,84
7	3,55	3,53	3,51	3,49	3,48	3,47	3,46	3,44	3,43	3,43	3,42	3,41
8	3,26	3,24	3,22	3,20	3,19	3,17	3,16	3,15	3,14	3,13	3,12	3,12
9	3,05	3,03	3,01	2,99	2,97	2,96	2,95	2,94	2,93	2,92	2,91	2,90
10	2,89	2,86	2,85	2,83	2,81	2,80	2,79	2,77	2,76	2,75	2,75	2,74
11	2,76	2,74	2,72	2,70	2,69	2,67	2,66	2,65	2,64	2,63	2,62	2,61
12	2,66	2,64	2,62	2,60	2,58	2,57	2,56	2,54	2,53	2,52	2,51	2,51
13	2,58	2,55	2,53	2,51	2,50	2,48	2,47	2,46	2,45	2,44	2,43	2,42
14	2,51	2,48	2,46	2,44	2,43	2,41	2,40	2,39	2,38	2,37	2,36	2,35
15	2,45	2,42	2,40	2,38	2,37	2,35	2,34	2,33	2,32	2,31	2,30	2,29
16	2,40	2,37	2,35	2,33	2,32	2,30	2,29	2,28	2,26	2,25	2,24	2,24
17	2,35	2,33	2,31	2,29	2,27	2,26	2,24	2,23	2,22	2,21	2,20	2,19
18	2,31	2,29	2,27	2,25	2,23	2,22	2,20	2,19	2,18	2,17	2,16	2,15
19	2,28	2,26	2,23	2,21	2,20	2,18	2,17	2,16	2,14	2,13	2,12	2,11
20	2,25	2,22	2,20	2,18	2,17	2,15	2,14	2,12	2,11	2,10	2,09	2,08
21	2,22	2,20	2,18	2,16	2,14	2,12	2,11	2,10	2,08	2,07	2,06	2,05
22	2,20	2,17	2,15	2,13	2,11	2,10	2,08	2,07	2,06	2,05	2,04	2,03
23	2,18	2,15	2,13	2,11	2,09	2,08	2,06	2,05	2,04	2,02	2,01	2,01
24	2,15	2,13	2,11	2,09	2,07	2,05	2,04	2,03	2,01	2,00	1,99	1,98
25	2,14	2,11	2,09	2,07	2,05	2,04	2,02	2,01	2,00	1,98	1,97	1,96
26	2,12	2,09	2,07	2,05	2,03	2,02	2,00	1,99	1,98	1,97	1,96	1,95
27	2,10	2,08	2,06	2,04	2,02	2,00	1,99	1,97	1,96	1,95	1,94	1,93
28	2,09	2,06	2,04	2,02	2,00	1,99	1,97	1,96	1,95	1,93	1,92	1,91
29	2,08	2,05	2,03	2,01	1,99	1,97	1,96	1,94	1,93	1,92	1,91	1,90
30	2,06	2,04	2,01	1,99	1,98	1,96	1,95	1,93	1,92	1,91	1,90	1,89
31	2,05	2,03	2,00	1,98	1,96	1,95	1,93	1,92	1,91	1,90	1,88	1,88
32	2,04	2,01	1,99	1,97	1,95	1,94	1,92	1,91	1,90	1,88	1,87	1,86
33	2,03	2,00	1,98	1,96	1,94	1,93	1,91	1,90	1,89	1,87	1,86	1,85
34	2,02	1,99	1,97	1,95	1,93	1,92	1,90	1,89	1,88	1,86	1,85	1,84
35	2,01	1,99	1,96	1,94	1,92	1,91	1,89	1,88	1,87	1,85	1,84	1,83
36	2,00	1,98	1,95	1,93	1,92	1,90	1,88	1,87	1,86	1,85	1,83	1,82
37	2,00	1,97	1,95	1,93	1,91	1,89	1,88	1,86	1,85	1,84	1,83	1,82
38	1,99	1,96	1,94	1,92	1,90	1,88	1,87	1,85	1,84	1,83	1,82	1,81
39	1,98	1,95	1,93	1,91	1,89	1,88	1,86	1,85	1,83	1,82	1,81	1,80
40	1,97	1,95	1,92	1,90	1,89	1,87	1,85	1,84	1,83	1,81	1,80	1,79
41	1,97	1,94	1,92	1,90	1,88	1,86	1,85	1,83	1,82	1,81	1,80	1,79
42	1,96	1,94	1,91	1,89	1,87	1,86	1,84	1,83	1,81	1,80	1,79	1,78

Scanned with OKEN Scanner

continuarea tabelului distribuției Fisher

Grade de libertate numitor	Grade de libertare numărător (25 - 36) ($\alpha = 0,05$)											
	25	26	27	28	29	30	31	32	33	34	35	36
1	249,3	249,5	249,6	249,8	250	250,1	250,2	250,4	250,5	250,6	250,7	250,8
2	19,46	19,46	19,46	19,46	19,46	19,46	19,46	19,46	19,47	19,47	19,47	19,47
3	8,63	8,63	8,63	8,62	8,62	8,62	8,61	8,61	8,61	8,61	8,60	8,60
4	5,77	5,76	5,76	5,75	5,75	5,75	5,74	5,74	5,74	5,73	5,73	5,73
5	4,52	4,52	4,51	4,50	4,50	4,50	4,49	4,49	4,48	4,48	4,48	4,47
6	3,83	3,83	3,82	3,82	3,81	3,81	3,80	3,80	3,80	3,79	3,79	3,79
7	3,40	3,40	3,39	3,39	3,38	3,38	3,37	3,37	3,36	3,36	3,36	3,35
8	3,11	3,10	3,10	3,09	3,08	3,08	3,07	3,07	3,07	3,06	3,06	3,06
9	2,89	2,89	2,88	2,87	2,87	2,86	2,86	2,85	2,85	2,85	2,84	2,84
10	2,73	2,72	2,72	2,71	2,70	2,70	2,69	2,69	2,69	2,68	2,68	2,67
11	2,60	2,59	2,59	2,58	2,58	2,57	2,57	2,56	2,56	2,55	2,55	2,54
12	2,50	2,49	2,48	2,48	2,47	2,47	2,46	2,46	2,45	2,45	2,44	2,44
13	2,41	2,41	2,40	2,39	2,39	2,38	2,38	2,37	2,37	2,36	2,36	2,35
14	2,34	2,33	2,33	2,32	2,31	2,31	2,30	2,30	2,29	2,29	2,28	2,28
15	2,28	2,27	2,27	2,26	2,25	2,25	2,24	2,24	2,23	2,23	2,22	2,22
16	2,23	2,22	2,21	2,21	2,20	2,19	2,19	2,18	2,18	2,17	2,17	2,17
17	2,18	2,17	2,17	2,16	2,15	2,15	2,14	2,14	2,13	2,13	2,12	2,12
18	2,14	2,13	2,13	2,12	2,11	2,11	2,10	2,10	2,09	2,09	2,08	2,08
19	2,11	2,10	2,09	2,08	2,08	2,07	2,07	2,06	2,06	2,05	2,05	2,04
20	2,07	2,07	2,06	2,05	2,05	2,04	2,03	2,03	2,02	2,02	2,01	2,01
21	2,05	2,04	2,03	2,02	2,02	2,01	2,00	2,00	1,99	1,99	1,98	1,98
22	2,02	2,01	2,00	2,00	1,99	1,98	1,98	1,97	1,97	1,96	1,96	1,95
23	2,00	1,99	1,98	1,97	1,97	1,96	1,95	1,95	1,94	1,94	1,93	1,93
24	1,97	1,97	1,96	1,95	1,95	1,94	1,93	1,93	1,92	1,92	1,91	1,91
25	1,96	1,95	1,94	1,93	1,93	1,92	1,91	1,91	1,90	1,90	1,89	1,89
26	1,94	1,93	1,92	1,91	1,91	1,90	1,89	1,89	1,88	1,88	1,87	1,87
27	1,92	1,91	1,90	1,90	1,89	1,88	1,88	1,87	1,87	1,86	1,86	1,85
28	1,91	1,90	1,89	1,88	1,88	1,87	1,86	1,86	1,85	1,85	1,84	1,84
29	1,89	1,88	1,88	1,87	1,86	1,85	1,85	1,84	1,84	1,83	1,83	1,82
30	1,88	1,87	1,86	1,85	1,85	1,84	1,83	1,83	1,82	1,82	1,81	1,81
31	1,87	1,86	1,85	1,84	1,83	1,83	1,82	1,82	1,81	1,81	1,80	1,80
32	1,85	1,85	1,84	1,83	1,82	1,82	1,81	1,80	1,80	1,79	1,79	1,78
33	1,84	1,83	1,83	1,82	1,81	1,81	1,80	1,79	1,79	1,78	1,78	1,77
34	1,83	1,82	1,82	1,81	1,80	1,80	1,79	1,78	1,78	1,77	1,77	1,76
35	1,82	1,82	1,81	1,80	1,79	1,79	1,78	1,77	1,77	1,76	1,76	1,75
36	1,81	1,81	1,80	1,79	1,78	1,78	1,77	1,76	1,76	1,75	1,75	1,74
37	1,81	1,80	1,79	1,78	1,77	1,77	1,76	1,76	1,75	1,74	1,74	1,73
38	1,80	1,79	1,78	1,77	1,77	1,76	1,75	1,75	1,74	1,74	1,73	1,73
39	1,79	1,78	1,77	1,77	1,76	1,75	1,75	1,74	1,73	1,73	1,72	1,72
40	1,78	1,77	1,77	1,76	1,75	1,74	1,74	1,73	1,73	1,72	1,72	1,71
41	1,78	1,77	1,76	1,75	1,74	1,74	1,73	1,72	1,72	1,71	1,71	1,70
42	1,77	1,76	1,75	1,75	1,74	1,73	1,72	1,72	1,71	1,71	1,70	1,70

Scanned with OKEN Scanner

continuarea tabelului distribuției Fisher

Grade de libertate numitor	Grade de libertare numărător (37 - 48) ($\alpha = 0,05$)											
	37	38	39	40	41	42	43	44	45	46	47	48
1	250,9	251	251,1	251,1	251,2	251,3	251,4	251,4	251,5	251,6	251,6	251,7
2	19,47	19,47	19,47	19,47	19,47	19,47	19,47	19,47	19,47	19,47	19,47	19,47
3	8,60	8,60	8,60	8,59	8,59	8,59	8,59	8,59	8,59	8,59	8,58	8,58
4	5,72	5,72	5,72	5,72	5,71	5,71	5,71	5,71	5,71	5,71	5,70	5,70
5	4,47	4,47	4,47	4,46	4,46	4,46	4,46	4,46	4,45	4,45	4,45	4,45
6	3,78	3,78	3,78	3,77	3,77	3,77	3,77	3,76	3,76	3,76	3,76	3,76
7	3,35	3,35	3,34	3,34	3,34	3,34	3,33	3,33	3,33	3,33	3,32	3,32
8	3,05	3,05	3,05	3,04	3,04	3,04	3,04	3,03	3,03	3,03	3,03	3,02
9	2,84	2,83	2,83	2,83	2,82	2,82	2,82	2,82	2,81	2,81	2,81	2,81
10	2,67	2,67	2,66	2,66	2,66	2,66	2,65	2,65	2,65	2,65	2,64	2,64
11	2,54	2,54	2,53	2,53	2,53	2,53	2,52	2,52	2,52	2,52	2,51	2,51
12	2,44	2,43	2,43	2,43	2,42	2,42	2,42	2,41	2,41	2,41	2,41	2,41
13	2,35	2,35	2,34	2,34	2,34	2,33	2,33	2,33	2,33	2,32	2,32	2,32
14	2,28	2,27	2,27	2,27	2,26	2,26	2,26	2,25	2,25	2,25	2,25	2,24
15	2,21	2,21	2,21	2,20	2,20	2,20	2,20	2,19	2,19	2,19	2,18	2,18
16	2,16	2,16	2,15	2,15	2,15	2,14	2,14	2,14	2,14	2,13	2,13	2,13
17	2,11	2,11	2,11	2,10	2,10	2,10	2,09	2,09	2,09	2,09	2,08	2,08
18	2,07	2,07	2,07	2,06	2,06	2,06	2,05	2,05	2,05	2,05	2,04	2,04
19	2,04	2,03	2,03	2,03	2,02	2,02	2,02	2,01	2,01	2,01	2,01	2,00
20	2,01	2,00	2,00	1,99	1,99	1,99	1,98	1,98	1,98	1,98	1,97	1,97
21	1,98	1,97	1,97	1,96	1,96	1,96	1,95	1,95	1,95	1,95	1,94	1,94
22	1,95	1,95	1,94	1,94	1,93	1,93	1,93	1,93	1,92	1,92	1,92	1,91
23	1,93	1,92	1,92	1,91	1,91	1,91	1,90	1,90	1,90	1,90	1,89	1,89
24	1,90	1,90	1,90	1,89	1,89	1,89	1,88	1,88	1,88	1,87	1,87	1,87
25	1,88	1,88	1,88	1,87	1,87	1,86	1,86	1,86	1,86	1,85	1,85	1,85
26	1,87	1,86	1,86	1,85	1,85	1,85	1,84	1,84	1,84	1,83	1,83	1,83
27	1,85	1,84	1,84	1,84	1,83	1,83	1,83	1,82	1,82	1,82	1,81	1,81
28	1,83	1,83	1,82	1,82	1,82	1,81	1,81	1,81	1,80	1,80	1,80	1,79
29	1,82	1,81	1,81	1,81	1,80	1,80	1,79	1,79	1,79	1,79	1,78	1,78
30	1,80	1,80	1,80	1,79	1,79	1,78	1,78	1,78	1,77	1,77	1,77	1,77
31	1,79	1,79	1,78	1,78	1,78	1,77	1,77	1,76	1,76	1,76	1,76	1,75
32	1,78	1,78	1,77	1,77	1,76	1,76	1,76	1,75	1,75	1,75	1,74	1,74
33	1,77	1,76	1,76	1,76	1,75	1,75	1,74	1,74	1,74	1,74	1,73	1,73
34	1,76	1,75	1,75	1,75	1,74	1,74	1,73	1,73	1,73	1,72	1,72	1,72
35	1,75	1,74	1,74	1,74	1,73	1,73	1,72	1,72	1,72	1,71	1,71	1,71
36	1,74	1,73	1,73	1,73	1,72	1,72	1,71	1,71	1,71	1,70	1,70	1,70
37	1,73	1,73	1,72	1,72	1,71	1,71	1,71	1,70	1,70	1,70	1,69	1,69
38	1,72	1,72	1,71	1,71	1,70	1,70	1,70	1,69	1,69	1,69	1,68	1,68
39	1,71	1,71	1,70	1,70	1,70	1,69	1,69	1,69	1,68	1,68	1,68	1,67
40	1,71	1,70	1,70	1,69	1,69	1,69	1,68	1,68	1,67	1,67	1,67	1,67
41	1,70	1,69	1,69	1,69	1,68	1,68	1,67	1,67	1,67	1,66	1,66	1,66
42	1,69	1,69	1,68	1,68	1,67	1,67	1,67	1,66	1,66	1,66	1,65	1,65

Scanned with OKEN Scanner

BIBLIOGRAFIE

1. Azoicăi Doina, Boiculese Lucian, Pisică-Donose George, Noțiuni de metodologie epidemiologică și statistică medicală, Editura Edit Dan, Iași, 2001
2. Berk Kenneth, Carey Patrick, Data Analysis with Microsoft Excel, Brooks/Cole, Cengage Learning, USA, 2010
3. Biji Elena-Maria, Lilea Eugenia, Gogu Emilia, Bentoiu Claudia Gabriela, Statistică. Aplicații practice, Editura Universitară, București, 2017
4. Blair Clifford, Taylor Richard, Biostatistics for the Health Sciences, Pearson Prentice Hall, USA, 2007
5. Bland Martin, An Introduction to Medical Statistics, Oxford University Press, UK, 2015
6. Boiculese Lucian, Dimitriu Gabriel, Moscalu Mihaela, Elemente de Biostatistică – Analiza statistică a datelor biologice, Editura Pim, Iași, 2007
7. Bruce Peter, Bruce Andrew, Practical Statistics for Data Scientists, O'Reilly Media, Inc., USA, 2017
8. Caragea Nicoleta, Alexandru Ciprian, Statistică – concepte, tehnici și instrumente software, Editura Pro Universitaria, București, 2018
9. Chap T. Le, Lynn E. Enerly, Introductory Biostatistics (Second edition), John Wiley & Sons, USA, 2016
10. Chernick Michael, The Essentials of Biostatistics for Physicians, Nurses, and Clinicians, John Wiley & Sons, Inc., USA, 2011
11. Cronk Brian, How to use SPSS. A Step-By-Step Guide to Analysis and Interpretation, Routledge, Taylor & Francis Group, New York, 2018
12. Darren George, Mallery Paul, IBM SPSS Statistics 26. Step by Step, (Sixteenth edition), Routledge Taylor & Francis Group, USA, New York, 2020
13. Drugan Tudor, Bolboacă Sorana, Leucuța Daniel, Bondor Cosmina, Călinici Tudor, Văleanu Mădălina, Colosi Horațiu, Iancu Mihaela, Istrate Dan, Curs de Biostatistică Medicală, Editura Medicală Universitară "Iuliu Hațieganu", Cluj-Napoca, 2018
14. Faizi Nafis, Alvi Yasir, Biostatistics Manual for Health Research, Academic Press, Elsevier, 2023
15. Farnell Damian, Medeiros Mirra Renata, Teaching Biostatistics in Medicine and Allied Health Sciences, Springer, 2023
16. Feinstein Alvan, Principles of Medical Statistics, Taylor & Francis Ltd., New York, 2001
17. Field Andy, Discovering Statistics using SPSS (Third edition), SAGE Publications LTD, UK, London, 2009

Scanned with OKEN Scanner

Manualul de faţă reprezintă o contribuţie remarcabilă la literatura de specialitate din domeniul statisticii aplicate în ştiinţele medicale. Cu un echilibru rafinat între rigoarea teoretică şi claritatea expunerii, autorul reuşeşte să transforme un domeniu perceput adesea ca arid într-un instrument accesibil, logic şi deosebit de util pentru studenţi, cercetători şi cadre medicale.

Structura coerentă, exemplele practice relevante şi accentul pus pe interpretarea critică a rezultatelor statistice fac din această lucrare un reper în formarea academică şi profesională a oricărui specialist din domeniul sănătăţii. Această carte poartă amprenta unei gândiri solide, animată de pasiune şi ghidată de un profund simţ al responsabilităţii ştiinţifice.

În ansamblu, lucrarea se distinge ca un demers ştiinţific de înaltă ţinută, al cărei conţinut merită pe deplin a fi integrat printre lucrările de referinţă din domeniul biostatisticii româneşti.

Sebastian IONESCU

Scanned with OKEN Scanner